ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: συλλογή, αναγνώριση και κατηγοριοποίηση γεγονότων

Σπυριδούλα Σκορδίλη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Θεόδωρος Τζουραμάνης

Επίκουρος Καθηγητής, Επιβλέπων

Λαμία, 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: συλλογή, αναγνώριση και κατηγοριοποίηση γεγονότων

Σπυριδούλα Σκορδίλη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Θεόδωρος Τζουραμάνης

Επίκουρος Καθηγητής

Λαμία, 2023

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 19/10/2023

Η Δηλούσα

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: συλλογή, αναγνώριση και κατηγοριοποίηση γεγονότων

Σπυριδούλα Σκορδίλη

Τριμελής Επιτροπή:

Θεόδωρος Τζουραμάνης, Επίκουρος Καθηγητής, Επιβλέπων

Κωνσταντίνος Δελήμπασης, Καθηγητής

Σωτήριος Τασουλής, Επίκουρος Καθηγητής

Ευχαριστίες

Θα ήθελα να εκφράσω τη βαθύτατη ευγνωμοσύνη μου στα άτομα που διαδραμάτισαν καθοριστικό ρόλο στην επιτυχή ολοκλήρωση της παρούσας πτυχιακής διατριβής.

Είμαι ευγνώμων στον επιβλέποντα καθηγητή μου, Θεόδωρο Τζουραμάνη, για την καθοδήγηση και την τεχνογνωσία του. Η αφοσίωσή του στην καλλιέργεια της πνευματικής μου ανάπτυξης, η διορατική του ανατροφοδότηση και η ικανότητά του να με κατευθύνει προς τη σωστή κατεύθυνση συνέβαλαν καθοριστικά στη διαμόρφωση αυτής της εργασίας.

Καθ' όλη τη διάρκεια αυτού του ταξιδιού, ο Θεόδωρος Τζουραμάνης δεν υπήρξε μόνο μέντορας αλλά και πηγή έμπνευσης. Είμαι επίσης ευγνώμων για τη συνεχή ενθάρρυνση, την εποικοδομητική ανατροφοδότηση και την ελευθερία που μου έδωσε να εξερευνήσω τις ιδέες μου και να προσεγγίσω αυτή την έρευνα με τρόπο που μου φάνηκε αυθεντικός. Η εμπιστοσύνη και η πίστη του στις ικανότητές μου με παρακίνησαν να διευρύνω τα όριά μου και να φτάσω σε νέα ύψη.

Επίσης, θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη και εκτίμηση στη συνερευνήτρια και φίλη μου, Ρόζμαρι Μάντζαρη, για την αδιάλειπτη αφοσίωση, τη συνεργασία και την υποστήριξή της καθ' όλη τη διάρκεια αυτής της ακαδημαϊκής προσπάθειας. Η διατριβή μου αποτελεί τμήμα ενός μεγαλύτερου έργου που εκπονήσαμε μαζί με τη Ρόζμαρι, δηλαδή της ανάπτυξης της πρώτης (ημι)αυτοματοποιημένης ελληνικής βάσης δεδομένων παγκόσμιων ειδησεογραφικών γεγονότων. Η παρούσα διατριβή αντιπροσωπεύει, λοιπόν, όχι μόνο τις δικές μου προσπάθειες αλλά και τη συλλογική σκληρή δουλειά και δέσμευση που έφερε στο τραπέζι και η Ρόζμαρι.

Η Ρόζμαρι και εγώ ξεκινήσαμε αυτό το ταξίδι μαζί, αντιμετωπίζοντας τις προκλήσεις και τους θριάμβους ως ομάδα. Οι μοναδικές γνώσεις μας, η ακούραστη προσπάθειά μας και τα εποικοδομητικά μας σχόλια ήταν ανεκτίμητα στη διαμόρφωση του περιεχομένου και της κατεύθυνσης αυτής της έρευνας. Ο κοινός μας ενθουσιασμός για το θέμα και οι αμέτρητες ώρες που περάσαμε δουλεύοντας δίπλα-δίπλα συνέβαλαν καθοριστικά στην υλοποίηση αυτού του έργου. Έτσι, η συνεργασία αυτή δεν εμπλούτισε μόνο την ακαδημαϊκή ποιότητα της παρούσας διατριβής, αλλά συνέβαλε επίσης σημαντικά στην προσωπική και επαγγελματική μου ανάπτυξη.

Αυτό το επίτευγμα είναι τόσο της Ρόζμαρι όσο και δικό μου, και είναι τιμή μου που είχα το προνόμιο να συνεργαστώ με έναν τόσο αφοσιωμένο και αξιόπιστο συνεργάτη. Η υποστήριξη και η συντροφιά μας ήταν ανεκτίμητη και είμαι πραγματικά τυχερή που μοιράστηκα αυτή την εμπειρία μαζί της.

Φυσικά, θα ήθελα να ευχαριστήσω τα άτομα του στενού μου κύκλου, συγγενείς ή μη, για την πίστη τους σε εμένα αλλά και την ψυχολογική τους στήριξη που με βοήθησε να επιτύχω μέχρι τέλους τον στόχο μου.

Περιεχόμενα

Μέρος Α: Θεωρητικό υπόβαθρο και σχετικές εργασίες

1 Εισαγωγή – Στόχος της πτυχιακής εργασίας

1.1 Εισαγωγή

Με το πέρας των χρόνων και την εξέλιξη της τεχνολογίας, ο Παγκόσμιος Ιστός αποτελεί σημαντικό εργαλείο ενημέρωσης και παρέχει το μεγαλύτερο σύνολο πληροφοριών παγκοσμίως που συνεχίζει να αναπτύσσεται με αυξανόμενο ρυθμό [A02]. Επομένως, τα δεδομένα στον Ιστό δεν είναι στατικά αλλά αντιθέτως, παράγονται και τροποποιούνται με εξαιρετική ταχύτητα [KJS20]. Σε συνδυασμό με τη δυνατότητα εύκολης πρόσβασης στα δεδομένα του Παγκόσμιου Ιστού και του μηδαμινού κόστους, το διαδίκτυο έχει αναδειχθεί ως αναπόσπαστο κομμάτι της καθημερινότητας του ανθρώπου.

Τα δεδομένα που διατίθενται στον Παγκόσμιο Ιστό χαρακτηρίζονται από τεράστιο όγκο, ο οποίος μετράται σε Zettabytes, δηλαδή δισεκατομμύρια gigabytes [KJS20]. Αυτό, έχει ως αποτέλεσμα, να έχουμε μεγάλη ποικιλομορφία δεδομένων, που βασίζονται σε μια ποικιλία τεχνολογικών και κανονιστικών προτύπων, τα οποία μπορεί να αποτελούν δομημένα, ημιδομημένα και μη δομημένα ποσοτικά και ποιοτικά δεδομένα. Μπορούν να εμφανιστούν με τη μορφή ιστοσελίδων, πινάκων HTML, βάσεων δεδομένων Ιστού, email, tweets, αναρτήσεων ιστολογίου, φωτογραφιών, βίντεο κ.λπ., τα οποία είναι ως επί το πλείστων αδόμητα [WKGS19].

Η ανάγκη για ενημέρωση οδηγεί ένα μεγάλο ποσοστό χρηστών να ανατρέχει καθημερινά σε μεγάλα διεθνή και Ελληνικά ειδησεογραφικά site, για να αναζητήσει τοπικές, εθνικές ή διεθνείς ειδήσεις. Οι ειδήσεις δημοσιεύονται με ταχύ ρυθμό καθημερινώς, γεγονός που δυσκολεύει τους ειδικούς του ειδησεογραφικού τομέα καθώς και τους πολίτες στην αναζήτηση πληροφοριών που έχουν δημοσιευθεί παλιότερα [A22]. Το πρόβλημα της αναζήτησης πληροφοριών σε διαδικτυακά άρθρα ειδήσεων δεν είναι η έλλειψή τους αλλά ο κατακλυσμός από αυτά. Αυτό δημιουργεί τεράστιες προκλήσεις στην επεξεργασία των διαδικτυακών ροών ειδήσεων, όπως είναι ο προσδιορισμός ποιου άρθρου ειδήσεων είναι σημαντικό, ο καθορισμός της ποιότητας κάθε άρθρου ειδήσεων και το φιλτράρισμα άσχετων και περιττών πληροφοριών.

Όπως ήδη αναφέρθηκε, η πρόσβαση σε διαδικτυακές ειδήσεις είναι μια από τις πλέον συνήθεις δραστηριότητες των χρηστών του Διαδικτύου. Συνεπώς, η εξάλειψη των περιττών ενημερωτικών ειδήσεων μπορεί να βοηθήσει στην εξοικονόμηση χρόνου όσο αφορά τον εντοπισμό χρήσιμων πληροφοριών και την αποφυγή της διαδικασίας φιλτραρίσματος αναπαραγόμενων πληροφοριών [GN06]. Επομένως, η κατασκευή μιας ολοκληρωμένης βάσης δεδομένων γεγονότων, με αυτόματη εξαγωγή από τον Ιστό [WKGS19], χρησιμεύει στη διευκόλυνση της διαχείρισης των ειδησεογραφικών δεδομένων.

1.2 Στόχος πτυχιακής

Η παρούσα πτυχιακή εργασία είναι ένα τμήμα ενός ευρύτερου έργου που στόχο έχει την δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων και υλοποιήθηκε από εμένα και την συνερευνήτριά μου Ρόζμαρι Μάντζαρη. Κατά συνέπεια, ένα τμήμα της τεχνικής περιγραφής του ευρύτερου αυτού έργου παρουσιάζεται στην εργασία μου και ένα δεύτερο τμήμα στην εργασία [M23]. Συγκεκριμένα στην παρούσα εργασία, παρουσιάζεται η υλοποίηση της συλλογής ειδήσεων, και της αναγνώρισης και κατηγοριοποίησης γεγονότων από πέντε διαφορετικές ειδησεογραφικές ιστοσελίδες πανελλήνιας εμβέλειας που παρέχουν κάλυψη γεγονότων από όλο τον κόσμο. Η ανάπτυξη του εν λόγω συστήματος πραγματοποιήθηκε σε γλώσσα προγραμματισμού Python καθώς έγινε χρήση και της MySQL για την αποθήκευση των γεγονότων σε βάση δεδομένων. Ειδικότερα, θα αναφερθούν τεχνικές και υλοποιήσεις της συλλογής, της προεπεξεργασίας, της κατηγοριοποίησης και της αποθήκευσης των γεγονότων στις επόμενες ενότητες. Στη συνέχεια, η Ρ. Μάντζαρη στην εργασία της [M23] παρουσιάζει μια τεχνική εντοπισμού διπλοτύπων και σχεδόν διπλοτύπων ειδήσεων που πραγματοποιεί πολλαπλούς ελέγχους τόσο κατά την διάρκεια εκτέλεσης του κώδικα συλλογής και ανάλυσης των ειδήσεων, όσο και κατά την αναγνώριση και αποθήκευση των γεγονότων στη βάση δεδομένων. Επίσης, στην εργασία [M23] παρουσιάζεται η υλοποίηση μιας λειτουργίας παρακολούθησης της εξέλιξης σημαντικών γεγονότων στο πέρασμα του χρόνου. Δημιουργήθηκε ένα άρτιο και ολοκληρωμένο αποτέλεσμα που μπορεί να χρησιμοποιηθεί και από τον οποιοδήποτε ενδιαφερόμενο ελληνόγλωσσο αναγνώστη που επιθυμεί να λάβει γνώση για τις εξελίξεις στην Ελλάδα και στο εξωτερικό αλλά και από κάθε ερευνητή που επιθυμεί να πραγματοποιήσει έρευνα πάνω σε ειδήσεις και γεγονότα.

1.3 Η έννοια του γεγονότος

Τα γεγονότα είναι αναπόσπαστο κομμάτι της ζωής. Για την βαθύτερη κατανόηση, κρίνεται απαραίτητος ο ορισμός του γεγονότος σύμφωνα με τις οποίες έχει δημιουργηθεί η παρούσα πτυχιακή εργασία. Βέβαια, δεν έχει καθοριστεί ένας συγκεκριμένος ορισμός για τη λέξη «γεγονός», όμως υπάρχουν διάφορες ερμηνείες με βάση τη βιβλιογραφία. Μερικές από αυτές τις ερμηνείες είναι:

> «Κάτι που συνέβη σε μια δεδομένη χρονική στιγμή ή σε μια συγκεκριμένη περίσταση· (συμβάν): ~ σπάνιο / συνηθισμένο / τυχαίο.»
>
> «Κάτι που συμβαίνει ή μπορεί να συμβεί.»
>
> «Κάτι με ιδιαίτερη σημασία και σπουδαιότητα.»
>
> «Κάτι αναμφισβήτητο.»
>
> «Κάτι συγκεκριμένο, πάντα αξιοσημείωτο.»

Γίνεται ταξινόμηση συχνά στα γεγονότα, θεωρώντας πολλά μεμονωμένα συμβάντα μαζί ως στιγμιότυπα κάποιας κατηγορίας ή τύπου συμβάντος, στα οποία δίνεται μία ονομασία, για παράδειγμα «αεροπορικά ατυχήματα». Κάθε τύπος συμβάντος δημιουργεί το δικό του συγκεκριμένο σύνολο προβλημάτων, αλλά μπορούν να αναγνωριστούν δύο ως θεμελιώδη: «Πώς ορίζεται ο τύπος συμβάντος;» και "Πώς μπορούμε να ανιχνεύσουμε πότε έχει συμβεί ένα συμβάν αυτού του τύπου;". Αυτές οι δύο ερωτήσεις είναι σαφώς στενά αλληλένδετες, αλλά δεν είναι εμφατικά η ίδια ερώτηση [GA02].

Ιστορικό γεγονός είναι αυτό που συνέβη ή συμβαίνει στην πραγματικότητα. Κάθε προσπάθεια ερμηνείας και ανάλυσής του, διατύπωσης κρίσης και έκφρασης γνώμης, είναι άποψη. Βέβαια, δεν είναι όλες οι απόψεις ισότιμες ούτε έχουν όλες την ίδια εγκυρότητα. Εύλογη είναι η διάκριση μεταξύ ιστορικού γεγονότος και απλού συμβάντος. Αν, για παράδειγμα, πραγματοποιηθεί μία αεροπορική πτήση, αυτό το δρομολόγιο είναι ένα απλό συμβάν. Ένα αεροπορικό δυστύχημα όμως, είναι ένα ιστορικό γεγονός, αφού έχει ευρύτερη επίδραση.

Τα συμβάντα μπορούν να απασχολήσουν τον ιστορικό ως σειρές αν μελετά νοοτροπίες ή ενδιαφέρεται για την κοινωνική ιστορία. Τα γεγονότα γίνονται αναντίρρητα αποδεκτά, καθώς υπάρχουν έξω από τον άνθρωπο, ανεξάρτητα από τη

γνώμη και τα συναισθήματά του ατόμου. Όσο και αν αυτή η διάκριση εκ πρώτης όψεως φαίνεται απλή, δεν είναι πάντοτε αυτονόητη. Τα γεγονότα πολύ συχνά είναι με κάποιον τρόπο στενά συνδεδεμένα με τις αξίες και τις πεποιθήσεις, τα συμφέροντά, με πατροπαράδοτες συνήθειες και στερεοτυπικές αντιλήψεις, με τη θρησκευτική πίστη και με τον τρόπο που ερμηνεύουμε τα πράγματα [IBE04]. Για τον λόγο αυτό ορίστηκε στα πλαίσια της εργασίας, πως για να χαρακτηριστεί μια είδηση ως γεγονός θα πρέπει να απαντά στα εξής ερωτήματα:

- ΠΟΙΟΣ; : όπου εκφράζει τον κύριο χαρακτήρα που αναφέρεται στο γεγονός.
- ΠΟΤΕ; : όπου δηλώνει το χρονικό πλαίσιο είτε της δημοσίευσης του γεγονότος είτε της πραγματοποίησής του.
- ΠΟΥ; : όπου προβάλλει την γεωγραφική θέση που έλαβε χώρα το γεγονός.
- ΤΙ; : όπου περιγράφει το πραγματικό γεγονός

Με τον τρόπο αυτό περιγράφεται πλήρως ένα γεγονός το οποίο είναι σημαντικό να συλλεχθεί [FA07] [FA09].

2 Πληροφοριακά συστήματα καταγραφής ειδησεογραφικών γεγονότων σε πραγματικό χρόνο

2.1 Εισαγωγή

Οι βάσεις δεδομένων διαδραματίζουν σημαντικό ρόλο στην ανάλυση και τη διαχείριση περιστατικών, είτε πρόκειται για τραγικά γεγονότα, όπως αυτοκτονίες, είτε για πολιτικά σημαντικά γεγονότα. Συλλέγοντας, οργανώνοντας και ερμηνεύοντας δεδομένα, οι βάσεις δεδομένων παρέχουν ένα δομημένο πλαίσιο για την κατανόηση και την αντιμετώπιση των προκλήσεων που συνδέονται με τέτοια γεγονότα.

Η κεντρική διαχείριση των δεδομένων αποτελεί σημαντικό πλεονέκτημα των βάσεων δεδομένων. Λειτουργούν ως πολύτιμα αποθετήρια για την αποθήκευση και την ασφαλή πρόσβαση σε πληροφορίες που σχετίζονται με συμβάντα. Συγκεντρώνοντας τα δεδομένα, διευκολύνουν την πρόσβαση και παρέχουν μια πλήρη εικόνα του συμβάντος. Η αποτελεσματική πρόσβαση στα δεδομένα είναι ένα άλλο πλεονέκτημα των βάσεων δεδομένων. Μετά από ένα περιστατικό, η έγκαιρη πρόσβαση σε σχετικές πληροφορίες είναι ζωτικής σημασίας. Οι βάσεις δεδομένων διευκολύνουν τις γρήγορες και στοχευμένες αναζητήσεις αλλά και βοηθούν τους ερευνητές να λάβουν πληροφορίες και τα απαραίτητα μέτρα.

Τα εργαλεία ανάλυσης μπορούν να βοηθήσουν στον εντοπισμό μοτίβων, σχέσεων και να παρέχουν βαθύτερη κατανόηση των υποκείμενων παραγόντων και των πιθανών αιτιών. Η ανάλυση αυτή μπορεί να δώσει πληροφορίες για προληπτικά μέτρα, ανάπτυξη πολιτικής και παρεμβάσεις. Η βάση δεδομένων υποστηρίζει επίσης, τη συνεργασία και την ανταλλαγή πληροφοριών μεταξύ των ενδιαφερομένων μερών που εμπλέκονται στην ανάλυση και τη διαχείριση συμβάντων. Επιπλέον, η βάση δεδομένων διατηρεί ακριβή ιστορικά αρχεία των συμβάντων. Αυτά τα ιστορικά δεδομένα αποτελούν πολύτιμο πόρο για μελλοντικές αναφορές, έρευνα και ανάπτυξη πολιτικής.

Στις επόμενες υποενότητες εξετάζονται διάφοροι τύποι βάσεων δεδομένων και οι εφαρμογές τους στην ανάλυση και διαχείριση συμβάντων. Θα παρουσιαστούν ορισμένα γνωστά χαρακτηριστικά τους και θα τονιστεί η σημασία τους στον επιστημονικό και ερευνητικό τομέα.

2.2 Global Database of Events, Language and Tone (GDELT)

Η Παγκόσμια Βάση Δεδομένων Γεγονότων, Γλώσσας και Τόνου (GDELT) είναι ένα δημοσίως διαθέσιμο σύνολο δεδομένων που παρακολουθεί και αναλύει παγκόσμια γεγονότα και ειδήσεις. Στόχος του είναι να καταγράψει και να κατανοήσει την περίπλοκη αλληλεπίδραση μεταξύ γεγονότων και της κάλυψης τους από τα μέσα ενημέρωσης.

Το GDELT περιλαμβάνει ένα ευρύ φάσμα πληροφοριών, συμπεριλαμβανομένων άρθρων ειδήσεων, αναρτήσεων στα μέσα κοινωνικής δικτύωσης και άλλων πηγών. Χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και μηχανικής μάθησης για την εξαγωγή και ανάλυση δεδομένων για γεγονότα, τοποθεσίες, εμπλεκόμενες οντότητες, θέματα και συναισθήματα.

Τα βασικά χαρακτηριστικά του GDELT περιλαμβάνουν:

1. *Δεδομένα συμβάντων*: Το GDELT συλλέγει και κατηγοριοποιεί συμβάντα από πηγές ειδήσεων παγκοσμίως. Αυτά τα γεγονότα μπορεί να περιλαμβάνουν πολιτικές συγκρούσεις, διαμαρτυρίες, διπλωματικές δραστηριότητες, ανθρωπιστικές κρίσεις και διάφορα άλλα είδη γεγονότων που διαμορφώνουν τις παγκόσμιες υποθέσεις.
2. *Γεωγραφική κάλυψη*: Το GDELT καλύπτει γεγονότα από όλο τον κόσμο, παρέχοντας μια ευρεία γεωγραφική αναπαράσταση και επιτρέποντας την ανάλυση σε τοπική, περιφερειακή και διεθνή κλίμακα.
3. *Χρονική κάλυψη*: Το σύνολο δεδομένων καλύπτει μια σημαντική χρονική περίοδο, ξεκινώντας από το 1979 έως σήμερα, επιτρέποντας την ιστορική ανάλυση και τον εντοπισμό μακροπρόθεσμων τάσεων και προτύπων.
4. *Ανάλυση γλώσσας και συναισθήματος*: Το GDELT χρησιμοποιεί τεχνικές NLP για την εξαγωγή και ανάλυση πληροφοριών που σχετίζονται με τη γλώσσα και τα συναισθήματα που εκφράζονται σε περιεχόμενο πολυμέσων. Αυτό δίνει τη δυνατότητα στους ερευνητές να μελετήσουν τον τόνο, το πλαίσιο και τον δημόσιο λόγο γύρω από συγκεκριμένα γεγονότα ή θέματα.
5. *Ανοιχτή πρόσβαση και εργαλεία ανάλυσης*: Το GDELT είναι ελεύθερα προσβάσιμο και παρέχει διάφορα εργαλεία και API που επιτρέπουν στους ερευνητές να εξερευνήσουν, να οπτικοποιήσουν και να αναλύσουν τα

δεδομένα. Αυτοί οι πόροι διευκολύνουν τη σε βάθος ανάλυση και ερμηνεία του συνόλου δεδομένων.

Το GDELT έχει χρησιμοποιηθεί από ερευνητές, δημοσιογράφους, υπεύθυνους χάραξης πολιτικής και οργανισμούς σε ένα ευρύ φάσμα πεδίων, συμπεριλαμβανομένων των πολιτικών επιστημών, των διεθνών σχέσεων, της ανάλυσης συγκρούσεων, των μελετών μέσων ενημέρωσης και της δημοσιογραφίας δεδομένων. Προσφέρει τη δυνατότητα κατανόησης των παγκόσμιων γεγονότων και των αφηγήσεων των μέσων ενημέρωσης σε μεγάλη κλίμακα. [LS13]

Εικόνα 1: Παράδειγμα ανάλυσης GDELT: Ο διαδραστικός χάρτης Foreign Policy του παγκόσμιου εγκλήματος στην άγρια ζωή (wildlife), όπως παράνομο κυνήγι. Συγκεκριμένα, αναφέρεται στις παγκόσμιες ειδήσεις από τον Φεβρουάριο έως τον Ιούνιο του 2015. Κάθε κουκκίδα αντιπροσωπεύει μια τοποθεσία που αναφέρεται σε ένα ή περισσότερα άρθρα που σχετίζονται με οποιαδήποτε μορφή εγκλήματος στην άγρια ζωή. Φωτογραφία: Kalev Leetaru.

2.3 Mass Mobilization Protest Dataset (MMPD)

Το Mass Mobilization Protest Dataset (MMPD) είναι ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων που παρέχει πληροφορίες για μαζικές διαδηλώσεις σε όλο τον κόσμο. Επικεντρώνεται συγκεκριμένα σε γεγονότα όπου ένας σημαντικός αριθμός ατόμων συμμετέχει σε συλλογικές ενέργειες για να εκφράσει τα παράπονά του ή να υποστηρίξει την κοινωνική ή πολιτική αλλαγή.

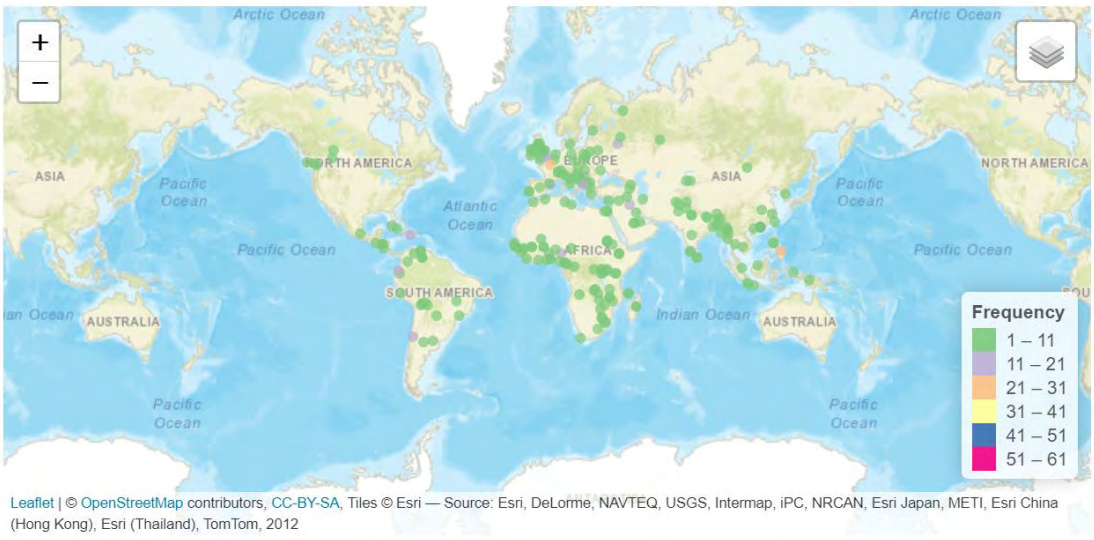
Το MMPD έχει σχεδιαστεί για να συλλέγει πληροφορίες σχετικά με διάφορες πτυχές των εκδηλώσεων διαμαρτυρίας, συμπεριλαμβανομένης της τοποθεσίας, της ημερομηνίας, των συμμετεχόντων, των στόχων, των τακτικών και των αποτελεσμάτων.

Καλύπτει ένα ευρύ φάσμα δραστηριοτήτων που σχετίζονται με διαμαρτυρίες, όπως πορείες, συγκεντρώσεις, απεργίες και άλλες μορφές μαζικής κινητοποίησης.

Τα βασικά χαρακτηριστικά του MMPD περιλαμβάνουν:

1. *Παγκόσμια κάλυψη:* Το σύνολο δεδομένων περιλαμβάνει διαμαρτυρίες από όλο τον κόσμο, παρέχοντας μια ευρεία γεωγραφική αναπαράσταση και επιτρέποντας συγκριτικές αναλύσεις μεταξύ χωρών και περιοχών.
2. *Λεπτομερείς πληροφορίες εκδήλωσης:* Το MMPD παρέχει ολοκληρωμένα δεδομένα για κάθε εκδήλωση διαμαρτυρίας, συμπεριλαμβανομένων των ειδικών στόχων της διαμαρτυρίας, των τακτικών που χρησιμοποιούν οι συμμετέχοντες, του επιπέδου οργάνωσης και τυχόν σχετικών αποτελεσμάτων ή παραχωρήσεων που έχουν επιτευχθεί.
3. *Διαχρονική προοπτική:* Το σύνολο δεδομένων καλύπτει μια σημαντική χρονική περίοδο, που ξεκινά από τα τέλη της δεκαετίας του 1990 ή τις αρχές της δεκαετίας του 2000, ανάλογα με την έκδοση του συνόλου δεδομένων. Αυτό επιτρέπει την ανάλυση των τάσεων, των προτύπων και των αλλαγών στη δυναμική των διαμαρτυριών με την πάροδο του χρόνου.
4. *Αναλυτικές ικανότητες*: Οι ερευνητές μπορούν να χρησιμοποιήσουν το MMPD για να εξετάσουν διάφορες διαστάσεις των εκδηλώσεων διαμαρτυρίας, όπως τους παράγοντες που συμβάλλουν στην κινητοποίηση, τις στρατηγικές που εφαρμόζουν οι διοργανωτές διαμαρτυρίας και τις συνέπειες της μαζικής κινητοποίησης για πολιτική και κοινωνική αλλαγή.

Το MMPD έχει χρησιμοποιηθεί ευρέως από ερευνητές, πολιτικούς επιστήμονες και κοινωνιολόγους που ενδιαφέρονται για τη μελέτη των κοινωνικών κινημάτων, του πολιτικού ακτιβισμού και της δυναμικής των διαμαρτυριών. Αποτελεί μια πολύτιμη πηγή για την κατανόηση της φύσης και του αντίκτυπου των μαζικών διαμαρτυριών σε διαφορετικά πλαίσια. [MMPD]



Εικόνα 2: Παράδειγμα ανάλυσης MMPD: Κάθε κουκίδα αντιπροσωπεύει μια τοποθεσία στην οποία πραγματοποιήθηκε μαζική κινητοποίηση το έτος 2012.

2.4 Armed Conflict Location & Event Data (ACLED)

Το Armed Conflict Location & Event Data (ACLED) είναι ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων που παρέχει λεπτομερείς πληροφορίες για γεγονότα πολιτικής βίας και ένοπλων συγκρούσεων σε όλο τον κόσμο. Συλλέγει και αναλύει δεδομένα για ένα ευρύ φάσμα περιστατικών που σχετίζονται με συγκρούσεις, όπως βομβαρδισμοί, ταραχές, διαμαρτυρίες και άλλες μορφές βίας.

Εστιάζει στην παρακολούθηση και την τεκμηρίωση γεγονότων που σχετίζονται με πολιτική αστάθεια, ένοπλες συγκρούσεις και κοινωνικές αναταραχές. Καλύπτει τόσο διεθνείς όσο και εσωτερικές συγκρούσεις και περιλαμβάνει δεδομένα από ένα ευρύ φάσμα πηγών, όπως ρεπορτάζ ειδήσεων, δημοσιεύσεις ΜΚΟ (Μη Κερδοσκοπικών Οργανισμών) και τοπικές πηγές. Το σύνολο δεδομένων παρέχει πληροφορίες για την τοποθεσία, την ημερομηνία, τους εμπλεκόμενους φορείς, τους τύπους βίας και τους θανάτους που σχετίζονται με κάθε καταγεγραμμένο συμβάν.

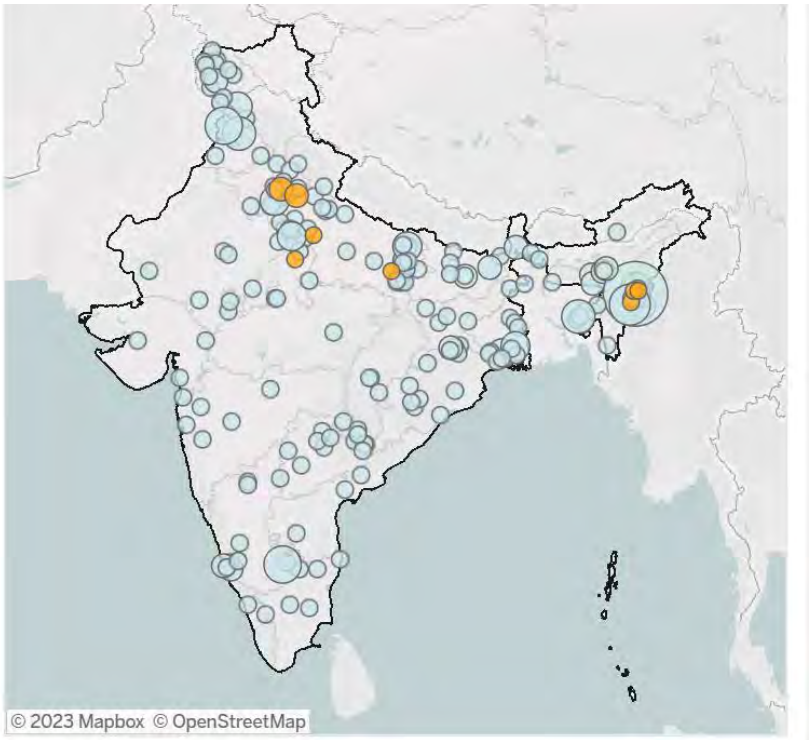
Τα βασικά χαρακτηριστικά του ACLED περιλαμβάνουν:

1. *Γεωγραφική κάλυψη*: Καλύπτει γεγονότα παγκοσμίως και παρέχει μια ευρεία γεωγραφική αναπαράσταση συγκρούσεων, συλλέγοντας δεδομένα από ηπείρους όπως η Αφρική, η Ασία, η Ευρώπη και η Αμερική.
2. *Χαρτογράφηση συμβάντος*: Το σύνολο δεδομένων παρέχει λεπτομερείς πληροφορίες για μεμονωμένα συμβάντα, επιτρέποντας ολοκληρωμένη

ανάλυση της δυναμικής και των προτύπων των συγκρούσεων. Περιλαμβάνει γεωχωρικές συντεταγμένες για τοποθεσίες συμβάντων, επιτρέποντας τη χαρτογράφηση και τη χωρική ανάλυση.

3. *Έγκαιρες ενημερώσεις*: Ενημερώνεται τακτικά για να ενσωματώνει νέα συμβάντα και αναθεωρήσεις σε υπάρχουσες καταχωρήσεις. Αυτό διασφαλίζει ότι το σύνολο δεδομένων αντικατοπτρίζει τις πιο πρόσφατες εξελίξεις και είναι σχετικό για συνεχή έρευνα και ανάλυση.
4. *Ανοιχτή πρόσβαση και εργαλεία ανάλυσης*: Είναι ελεύθερα προσβάσιμο στο κοινό και οι χρήστες μπορούν να κάνουν λήψη των δεδομένων για ερευνητικούς σκοπούς. Το ACLED παρέχει επίσης διαδικτυακά εργαλεία και επιλογές οπτικοποίησης για να διευκολύνει την εξερεύνηση και ανάλυση δεδομένων.

Το ACLED έχει χρησιμοποιηθεί ευρέως από ερευνητές, υπεύθυνους χάραξης πολιτικής και οργανισμούς που ασχολούνται με την ανάλυση συγκρούσεων, την ανθρωπιστική απόκριση και τις προσπάθειες οικοδόμησης της ειρήνης. Παρέχει πολύτιμες γνώσεις σχετικά με τη δυναμική των συγκρούσεων, τις τάσεις και τον αντίκτυπο της βίας στους πληγέντες πληθυσμούς. [ACLED]



Εικόνα 3: Παράδειγμα ανάλυσης του ACLED: Οι πορτοκαλί κουκίδες αντιπροσωπεύουν κοινωνικές αναταραχές που πραγματοποιήθηκαν τον Μάιο του 2023. Οι μπλε κουκίδες αντιπροσωπεύουν κινητοποιήσεις που πραγματοποιήθηκαν πριν από τον Μάιο του 2023.

2.5 The Database on Suicide Attacks (DSAT)

Το Database on Suicide Attacks (DSAT) περιέχει το σύνολο δεδομένων CPOST, και αποτελεί μια βάση δεδομένων που εστιάζει ειδικά στις επιθέσεις αυτοκτονίας. Το σύνολο δεδομένων CPOST καταγράφει λεπτομερείς πληροφορίες για διάφορες πτυχές των επιθέσεων αυτοκτονίας, συμπεριλαμβανομένης της ημερομηνίας, της τοποθεσίας, των χαρακτηριστικών του δράστη, του τύπου στόχου, της μεθόδου επίθεσης, των θυμάτων και άλλων σχετικών μεταβλητών. Καλύπτει ένα παγκόσμιο φάσμα επιθέσεων αυτοκτονίας και στοχεύει σε μια ολοκληρωμένη κατανόηση αυτής της συγκεκριμένης μορφής βίας.

Τα βασικά χαρακτηριστικά του CPOST περιλαμβάνουν:

1. *Λεπτομέρειες δράστη*: Το σύνολο δεδομένων περιλαμβάνει πληροφορίες για τους δράστες που εμπλέκονται στις επιθέσεις αυτοκτονίας, όπως η σχέση τους, τα κίνητρά τους, τα δημογραφικά στοιχεία και οποιαδήποτε άλλη διαθέσιμη σχετική πληροφορία.
2. *Πληροφορίες στόχου*: Το CPOST καταγράφει τους τύπους-στόχους των επιθέσεων αυτοκτονίας, όπως στρατιωτικές εγκαταστάσεις, κυβερνητικά κτίρια, δημόσιους χώρους, θρησκευτικούς χώρους ή άλλους συγκεκριμένους στόχους.
3. *Μέθοδοι επίθεσης*: Το σύνολο δεδομένων παρέχει πληροφορίες σχετικά με τις μεθόδους που χρησιμοποιούνται σε επιθέσεις αυτοκτονίας, όπως η χρήση εκρηκτικών, οι επιθέσεις με όχημα ή άλλες τακτικές των δραστών.
4. *Δεδομένα ατυχημάτων*: Το CPOST καταγράφει τον αριθμό των θυμάτων που προκύπτουν από κάθε επίθεση αυτοκτονίας, συμπεριλαμβανομένων των θανάτων και των τραυματισμών.
5. *Συγκριτική ανάλυση*: Το σύνολο δεδομένων επιτρέπει τη συγκριτική ανάλυση σε διαφορετικές περιοχές και χρονικές περιόδους, επιτρέποντας στους ερευνητές να εντοπίσουν τάσεις, πρότυπα και παραλλαγές στις επιθέσεις αυτοκτονίας.

Το σύνολο δεδομένων CPOST για τις επιθέσεις αυτοκτονίας έχει χρησιμοποιηθεί ευρέως από ερευνητές, υπεύθυνους χάραξης πολιτικής και αναλυτές για τη μελέτη του φαινομένου των επιθέσεων αυτοκτονίας, τον εντοπισμό παραγόντων που συμβάλλουν

στην εμφάνισή τους και την ανάπτυξη στρατηγικών για την αντιμετώπιση και την πρόληψη τέτοιων πράξεων βίας. [PRC21]

Εικόνα 4: Παράδειγμα ανάλυσης CPOST: Κάθε κουκίδα αντιπροσωπεύει τρομοκρατική επίθεση. Ευρώπη 1970-2015.

2.6 The International Disasters Database (EM-DAT)

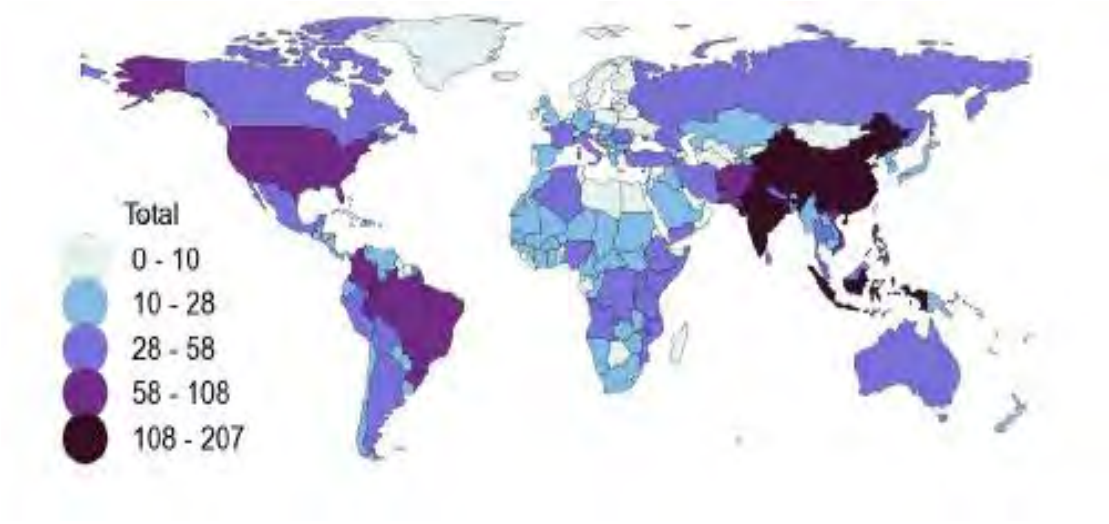
Η διεθνής βάση δεδομένων καταστροφών, κοινώς γνωστή ως EM-DAT, είναι μια βάση δεδομένων που παρέχει πληροφορίες για φυσικές και τεχνολογικές καταστροφές παγκοσμίως. Η EM-DAT εστιάζει στη συλλογή δεδομένων που σχετίζονται με καταστροφές που πληρούν συγκεκριμένα κριτήρια, συμπεριλαμβανομένης της απώλειας ζωών, του αριθμού των ανθρώπων που επηρεάζονται, της κήρυξης κατάστασης έκτακτης ανάγκης ή της έκκλησης για διεθνή βοήθεια. Η βάση δεδομένων περιλαμβάνει ένα ευρύ φάσμα καταστροφών, όπως σεισμούς, πλημμύρες, τυφώνες, ξηρασίες, βιομηχανικά ατυχήματα, επιδημίες και άλλα γεγονότα.

Μερικά βασικά χαρακτηριστικά της EM-DAT περιλαμβάνουν:

1. *Συλλογή δεδομένων:* Η EM-DAT συλλέγει πληροφορίες από διάφορες πηγές, όπως κυβερνητικές εκθέσεις, μέσα ενημέρωσης, επιστημονικές δημοσιεύσεις και διεθνείς οργανισμούς. Τα δεδομένα επικυρώνονται και διασταυρώνονται για να διασφαλιστεί η ακρίβεια και η αξιοπιστία.

2. *Λεπτομερείς πληροφορίες*: Η βάση δεδομένων παρέχει λεπτομερείς πληροφορίες για κάθε καταστροφή, λαμβάνοντας υπόψη την ημερομηνία και την τοποθεσία του συμβάντος, το είδος της καταστροφής, τον αριθμό των θανάτων, των τραυματισμών και των πληγέντων ατόμων, τις οικονομικές ζημιές και άλλες σχετικές λεπτομέρειες.
3. *Προσβασιμότητα*: Η EM-DAT είναι ελεύθερα προσβάσιμο στο κοινό. Οι χρήστες μπορούν να αναζητήσουν και να ανακτήσουν δεδομένα βάσει συγκεκριμένων κριτηρίων, όπως η τοποθεσία, ο τύπος της καταστροφής ή η χρονική περίοδος. Η βάση δεδομένων επιτρέπει επίσης στους χρήστες να δημιουργούν αναφορές και να κατεβάζουν δεδομένα για περαιτέρω ανάλυση.
4. *Στατιστική ανάλυση*: Η EM-DAT επιτρέπει στους ερευνητές και τους υπεύθυνους χάραξης πολιτικής να αναλύουν τις τάσεις, τα πρότυπα και τις επιπτώσεις των καταστροφών με την πάροδο του χρόνου. Τα δεδομένα μπορούν να χρησιμοποιηθούν για την αξιολόγηση των στρατηγικών αντιμετώπισης καταστροφών.

Η EM-DAT αναγνωρίζεται ευρέως ως αξιόπιστη και πολύτιμη πηγή για την κατανόηση της εμφάνισης των καταστροφών παγκοσμίως και των επιπτώσεών τους. Χρησιμοποιείται από ερευνητές, ανθρωπιστικούς οργανισμούς, υπεύθυνους χάραξης πολιτικής και ειδικούς που εμπλέκονται στη διαχείριση καταστροφών και τη μείωση του κινδύνου. Η EM-DAT ενημερώνει τακτικά τη βάση δεδομένων της με νέες πληροφορίες και αναθεωρήσεις για να εξασφαλίσει την πιο ακριβή αναπαράσταση των γεγονότων καταστροφών. [SM92]



Εικόνα 5:Παράδειγμα ανάλυσης EM-DAT: Πλημμύρες ανά χώρα από το 2000 μέχρι το 2022.

2.7 Επίλογος

Η χρήση των βάσεων δεδομένων στην ανάλυση και διαχείριση συμβάντων έχει αποδειχθεί ανεκτίμητη για την αντιμετώπιση των προκλήσεων που συνδέονται με ένα ευρύ φάσμα συμβάντων. Ολοκληρώνοντας την έρευνα σχετικά με τη σημασία των βάσεων δεδομένων στον τομέα αυτό, καθίσταται σαφές ότι διαδραματίζουν σημαντικό ρόλο στην κατανόηση και την πρόληψη των επιπτώσεων των συμβάντων.

Οι βάσεις δεδομένων επέτρεψαν στους αναλυτές να εμβαθύνουν στα δεδομένα των συμβάντων και να αποκαλύψουν μοτίβα, τάσεις και συνδέσεις που συμβάλλουν στη συνολική κατανόηση των συμβάντων. Η ικανότητα ανάλυσης και συσχέτισης των δεδομένων διευκόλυνε τον εντοπισμό των υποκείμενων παραγόντων, των πιθανών αιτιών και των συνεπειών. Οι πληροφορίες αυτές άνοιξαν το δρόμο για προληπτικές στρατηγικές, παρεμβάσεις πολιτικής και βελτιωμένες πρακτικές διαχείρισης συμβάντων.

Εκτός από την αναλυτική ικανότητα, η βάση δεδομένων διευκόλυνε την ανταλλαγή πληροφοριών και προώθησε τις συνεργατικές προσπάθειες παρέχοντας ελεγχόμενη πρόσβαση και ασφαλείς μηχανισμούς ανταλλαγής δεδομένων. Επιπλέον, τα ιστορικά αρχεία που είναι αποθηκευμένα στη βάση δεδομένων αποτελούν πολύτιμο πόρο για μελλοντικές αναφορές, έρευνα και χάραξη πολιτικής.

Είναι σημαντικό να συνεχιστεί η ανάπτυξη της ικανότητας των βάσεων δεδομένων για την ανάλυση και τη διαχείριση περιστατικών. Οι τεχνολογικές εξελίξεις, όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση, υπόσχονται να ενισχύσουν περαιτέρω τη δύναμη και την αποτελεσματικότητα των βάσεων δεδομένων στην κατανόηση των συμβάντων, την πρόβλεψη πιθανών κινδύνων και την καθοδήγηση της προληπτικής λήψης αποφάσεων.

Ως εκ τούτου, η σημασία των βάσεων δεδομένων στην ανάλυση και τη διαχείριση περιστατικών δεν μπορεί να υπερτονιστεί. Ο ρόλος των βάσεων δεδομένων στην υποστήριξη της συλλογής δεδομένων, της αποτελεσματικής πρόσβασης, της ανάλυσης και της συνεργασίας και της διαχείρισης ιστορικών αρχείων έχει αλλάξει τον τρόπο προσέγγισης των συμβάντων. Η αξιοποίηση της δύναμης τους μπορεί να δημιουργήσει ένα ασφαλέστερο περιβάλλον, να βελτιώσει τις δυνατότητες αντίδρασης και να προωθήσει μια πιο ενημερωμένη κοινωνία που θα μπορεί να αντιμετωπίσει τις μελλοντικές προκλήσεις.

3 Μέθοδοι συλλογής και καθαρισμού περιεχομένου άρθρων από ειδησεογραφικές πύλες στον παγκόσμιο ιστό

3.1 Εισαγωγή

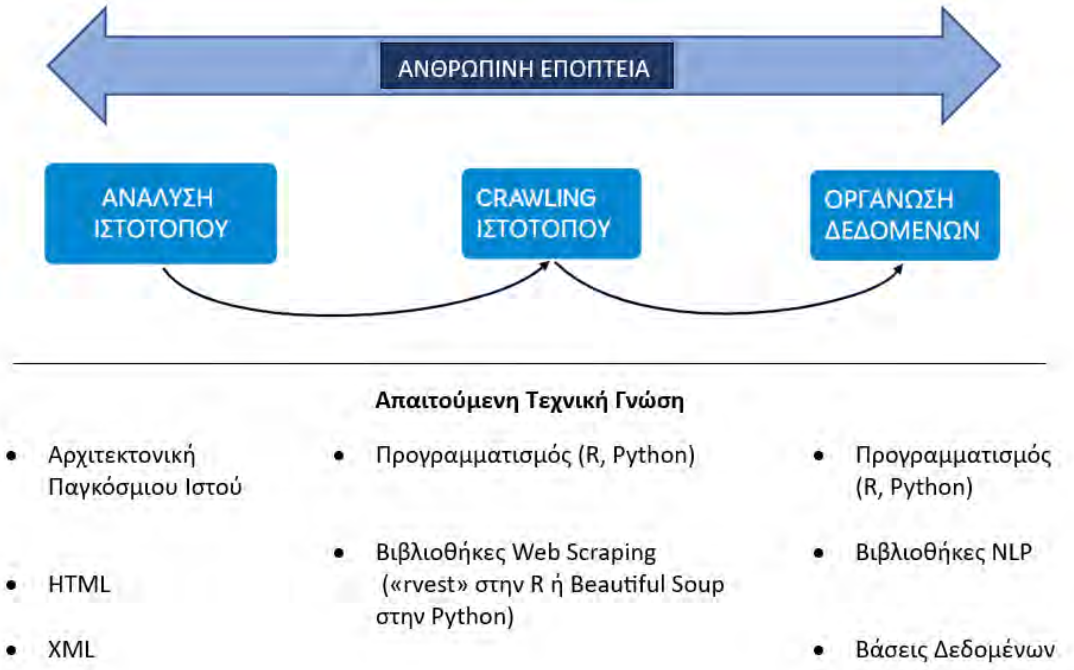
Η συλλογή, η οργάνωση και η ανάλυση των δεδομένων του Παγκόσμιου Ιστού είναι μια πολύπλοκη διαδικασία λόγω του μεγάλου όγκου, της ποικιλίας και της ταχύτητας ανανέωσης των δεδομένων. Για να αντιμετωπιστεί αυτή η πρόκληση, χρησιμοποιούνται τεχνολογίες και εργαλεία που αυτοματοποιούν τη συλλογή και την οργάνωση των δεδομένων Ιστού. Αυτή η διαδικασία αυτοματοποιημένης εξαγωγής και οργάνωσης δεδομένων από τον Παγκόσμιο Ιστό ονομάζεται απόξεση ιστού (Web Scraping ή Web Crawling).

Μετά τη συλλογή των δεδομένων, είναι απαραίτητο να υποβληθούν σε προεπεξεργασία, προκειμένου να καθαριστούν και να οργανωθούν κατάλληλα για την περαιτέρω ανάλυση. Η καλή προετοιμασία των δεδομένων είναι σημαντική για να επιτευχθεί μια αποτελεσματική ανάλυση, αφαιρώντας λάθη και ανακρίβειες που μπορεί να παρουσιάζονται κατά την επεξεργασία. Το διαδικτυακό κείμενο μπορεί να περιέχει ανεπιθύμητο περιεχόμενο, ακατέργαστες πληροφορίες και άλλα αδιάφορα δεδομένα όπως διαφημίσεις. Ανάλογα με τον όγκο των δεδομένων που εμπλέκονται, είναι αναγκαία η προγραμματιστική προσέγγιση για να εξοικονομηθεί χρόνος.

3.2 Μέθοδοι συλλογής άρθρων από ειδησεογραφικές πύλες στον παγκόσμιο ιστό

Δεδομένου του όγκου, της ποικιλίας και της ταχύτητας ανανέωσης των δεδομένων του Παγκόσμιου Ιστού, η συλλογή και η οργάνωση αυτών, δύσκολα μπορεί να γίνει χειροκίνητα. Εξαιτίας αυτού, συχνά χρησιμοποιούνται διάφορες τεχνολογίες και εργαλεία που αυτοματοποιούν ορισμένες ή όλες τις πτυχές της συλλογής και οργάνωσης των δεδομένων Ιστού. Η διαδικασία αυτόματης εξαγωγής και οργάνωσης δεδομένων με τη χρήση λογισμικού, με σκοπό την ανάλυση αυτών, ονομάζεται απόξεση ιστού (Web Scraping ή Web Crawling) [KJS20].

Η απόξεση ιστού είναι μια τεχνική για τη μετατροπή μη δομημένων δεδομένων Ιστού σε δομημένα δεδομένα που μπορούν να αποθηκευτούν και να αναλυθούν σε ένα κεντρικό υπολογιστικό φύλλο ή βάση δεδομένων [K21]. Η απόξεση ιστού αποτελείται από τις ακόλουθες τρεις κύριες φάσεις: ανάλυση ιστότοπου, ανίχνευση ιστότοπου και οργάνωση δεδομένων. Η υλοποίηση των τριών φάσεων απαιτεί τουλάχιστον μία γλώσσα προγραμματισμού που σχετίζεται με την Επιστήμη των Δεδομένων, όπως η R ή η Python, διότι υπάρχει διαθεσιμότητα βιβλιοθηκών, όπως το πακέτο «rvest» στην R ή τη βιβλιοθήκη Beautiful Soup στην Python, προκειμένου να γίνει η αυτόματη ανίχνευση και ανάλυση δεδομένων Ιστού. Η Επιστήμη των Δεδομένων είναι ένα διεπιστημονικό πεδίο σχετικά με επιστημονικές μεθόδους, διαδικασίες και συστήματα για την εξαγωγή γνώσεων.



Εικόνα 6: Φάσεις του Web Scraping - Web Crawling

Η πρώτη φάση, η ανάλυση του ιστότοπου, συνιστά τη γνώση της υποκείμενης δομής ενός ιστότοπου ή μίας διαδικτυακής βάσης δεδομένων προκειμένου να κατανοηθεί ο τρόπος με τον οποίο αποθηκεύονται τα απαραίτητα δεδομένα. Αυτό απαιτεί μια βασική κατανόηση της αρχιτεκτονικής του Παγκόσμιου Ιστού, των γλωσσών σήμανσης, όπως HTML, XML, και διαφόρων βάσεων δεδομένων, όπως η MySQL. Η δεύτερη φάση, η ανίχνευση ιστότοπου, περιλαμβάνει τον σχεδιασμό και την εκτέλεση ενός script που περιηγείται αυτόματα στον ιστότοπο και ανακτά τα απαραίτητα δεδομένα [KJS20]. Η τρίτη φάση, η οργάνωση των δεδομένων, μπορεί να

υλοποιηθεί με διαφορετικές μεθόδους ανά ερευνητή και εξαρτάται από τη πολυπλοκότητα και τη φύση του προβλήματος. Συχνά, οι τρεις φάσεις του Web Scraping δεν μπορούν να αυτοματοποιηθούν πλήρως και απαιτούν τουλάχιστον κάποιο βαθμό ανθρώπινης συμμετοχής.

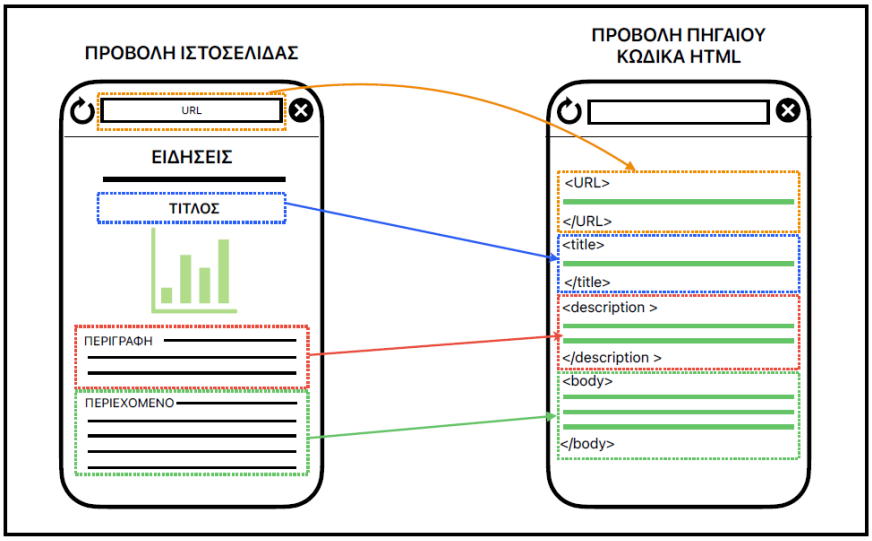
Η χρήση της απόξεσης ιστού μας επιτρέπει να συλλέγουμε δεδομένα, με το πρωτόκολλο HTTP που χρησιμοποιείται από προγράμματα περιήγησης ιστού, σε σχεδόν πραγματικό χρόνο από ιστότοπους. Έναντι της αντιγραφής και της συλλογής πληροφοριών χειροκίνητα από τους ερευνητές, αυτή τη λειτουργία προσφέρεται από το Web Scraping με μεγαλύτερη ακρίβεια και πολύ πιο γρήγορα από έναν άνθρωπο [K21] [SDN17]. Η απόξεση ιστού μπορεί να πραγματοποιηθεί με τις αρχιτεκτονικές ιστοτόπου HTML, XMR, RSS οι οποίες θα αναλυθούν στις παρακάτω υποενότητες.

3.2.1 Αρχιτεκτονική HTML:

Τα περισσότερα έγγραφα που είναι διαθέσιμα στον Παγκόσμιο Ιστό επιβεβαιώνουν τις προδιαγραφές HTML. Προορίζονται για ανάγνωση από τον άνθρωπο μέσω ενός προγράμματος περιήγησης ιστού και έτσι δομούνται σύμφωνα με ορισμένες κοινές συμβάσεις. Επομένως, το Εννοιολογικό Μοντέλο για HTML προτάθηκε στα τέλη του 1990 για να καταγράψει αυτόματα την ιεραρχική δομή στα έγγραφα του Ιστού [K21]. Ο Ιστός αναπτύχθηκε αρχικά ομοιογενώς γύρω από την HTML. Κατά την τεχνολογική εξέλιξη, ο Ιστός γινόταν όλο και πιο περίπλοκος [A02].

Ο web browser διαβάζει τα έγγραφα HTML και τα συνθέτει σε σελίδες. Δεν εμφανίζει τις ετικέτες, που είναι μέρος της αρχιτεκτονικής HTML, αλλά τις χρησιμοποιεί για τη παρουσίαση του περιεχομένου της σελίδας. Η HTML επιτρέπει την ενσωμάτωση εικόνων, βίντεο και άλλων διαδραστικών μέσων, ενώ παράλληλα παρέχει τις μεθόδους δημιουργίας δομημένων εγγράφων, προσδιορίζοντας σημαντικά δομικά στοιχεία για το κείμενο, όπως κεφαλίδες, παραγράφους, λίστες και άλλα.

Στην παρακάτω εικόνα, αναγράφονται η προβολή ιστοσελίδας και η προβολή του πηγαίου κώδικα HTML της αντίστοιχης ιστοσελίδας. Η προβολή ιστοσελίδας δείχνει τι βλέπει ένας αναγνώστης όταν διαβάζει ένα άρθρο ειδήσεων: τη διεύθυνση URL, τον τίτλο, την περιγραφή, το περιεχόμενο καθώς και την ημερομηνία και ώρα δημοσίευσης. Στην προβολή πηγαίου κώδικα HTML, υπάρχουν κατάλληλες ετικέτες οι οποίες αντιστοιχούν στα παραπάνω δεδομένα [AZHL22].



Εικόνα 7: Προβολή ιστοσελίδας και πηγαίου κώδικα HTML

3.2.2 Αρχιτεκτονική XML:

Η XML (Extensible Markup Language) είναι μια γλώσσα σήμανσης, δηλαδή, τα έγγραφα XML μπορούν να χωριστούν αυτόματα σε τμήματα εγγράφων χρησιμοποιώντας τη σήμανση. Ακόμα, είναι μια μορφή αρχείου κατάλληλη για την αποθήκευση, τη μετάδοση και την ανακατασκευή αδόμητων δεδομένων. Καθορίζει ένα σύνολο κανόνων για την κωδικοποίηση εγγράφων σε μορφή που είναι τόσο αναγνώσιμη από τον άνθρωπο όσο και αναγνώσιμη από μηχανή. Χρησιμοποιείται ευρέως ως τυπική μορφή εγγράφου σε πολλούς τομείς εφαρμογών. Η XML έχει τεθεί σε κοινή χρήση για την ανταλλαγή δεδομένων μέσω του Διαδικτύου. Έχουν αναπτυχθεί εκατοντάδες μορφές εγγράφων που χρησιμοποιούν σύνταξη XML, συμπεριλαμβανομένων των RSS , Atom , Office Open XML και άλλων [HKYU02]. Μπορεί να χρησιμοποιηθεί για την αναπαράσταση εγγράφων, αλλά και δομημένων ή ημιδομημένων δεδομένων [KJS20] και παντρεύει, κατά κάποιο τρόπο, τον κόσμο των εγγράφων με τον κόσμο της βάσης δεδομένων [HKYU02].

3.2.3 Αρχιτεκτονική RSS:

Τον Μάρτιο του 1999, ο Dan Libby δημιούργησε μια μέθοδο για τη διανομή πληροφοριών που ονομάζεται RSS2. Το RSS είναι μια σειρά αρχείων μορφοποιημένων XML για διανομή Ιστού που έχει γίνει πρότυπο για ειδησεογραφικές πύλες και έχει υιοθετηθεί ευρέως για την αναπαραγωγή άλλου τύπου πληροφοριών, όπως ιστολόγια, ιστοσελίδες καταστημάτων, ιστοσελίδες αγγελιών εργασίας και άλλων. Όπως έχει υποστηριχθεί στο Feedster3, «Ό,τι είναι έγκαιρο και πολύτιμο στον Ιστό θα είναι

διαθέσιμο ως ροή RSS». Το Feedster είναι μία ενιαία λίστα - ευρετήριο όλων των ροών RSS (RSS feed) / XML που έχει πρόσβαση και παρέχει μια μηχανή αναζήτησης για αυτό ακριβώς το περιεχόμενο. Η αναζήτηση αυτή μπορεί να αποθηκευτεί ως ροή RSS [GN06].

Με την ανάπτυξη του RSS, οι χρήστες του Διαδικτύου μπορούν να ανακτήσουν ένα αρχείο RSS που περιέχει όχι μόνο τους τίτλους, αλλά και μια σύντομη περιγραφή κάθε είδησης και τον σύνδεσμο που οδηγεί στη πηγή της είδησης σε μία σελίδα. Η τεχνολογία του RSS επιτρέπει στους χρήστες του Διαδικτύου να εγγραφούν σε τοποθεσίες Web που συνήθως προσθέτουν ή τροποποιούν περιεχόμενο τακτικά και γρήγορα. Για να χρησιμοποιήσουν αυτήν την τεχνολογία, οι υπεύθυνοι διαχείρισης των ιστοσελίδων δημιουργούν ή αποκτούν εξειδικευμένο λογισμικό, το οποίο είναι σε μορφή αναγνώσιμη από συστήματα XML [GN06].

Στην παρακάτω εικόνα παρατηρείται μια τυπική μορφή RSS feed.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<rss version="2.0">
  <channel>
    <title>NYT &gt; World Business</title>
        <item>
          <title>Russia and Ukraine Reach Compromise</title>
          <link>http://www.nytimes.com/2006/01/05/05ukraine.html</link>
          <description>The solution allowed both nations...</description>
          <author>ANDREW E. KRAMER</author>
          <pubDate>Thu, 05 Jan 2006 00:00:00 EDT</pubDate>
        </item>
    </title>
  </channel>
</rss>
```

Εικόνα 8: RSS feed. Πηγή εικόνας: [F06]

Ένα από τα βασικά στοιχεία σε ένα αρχείο RSS είναι το Channel, το οποίο περιέχει πολλά στοιχεία που περιγράφουν τις πληροφορίες που αναγράφονται στο αρχείο. Τα στοιχεία του καναλιού περιλαμβάνουν:

(i) Τον τίτλο (<title>), ο οποίος είναι παρόμοιος με την ετικέτα <title> σε ένα αρχείο HTML και χρησιμοποιείται για την αναγνώριση της ροής ειδήσεων RSS.

(ii) Τον σύνδεσμο (<link>), που είναι η διεύθυνση URL της τοποθεσίας Web από την οποία μπορεί να ανακτηθεί το αρχείο RSS.

(iii) Την περιγραφή (<description>), η οποία περιλαμβάνει μια πρόταση που περιγράφει εν συντομία τι αφορούν τα άρθρα ειδήσεων, που περιέχονται στο αρχείο, όπως πολιτική, οικονομία, αθλητισμός κ.λπ.

(iv) Στοιχείο (<item>), το οποίο είναι ολόκληρη η είδηση. Αντιμετωπίζουμε ένα στοιχείο ως μια πλειάδα μιας ροής ειδήσεων RSS, η οποία περιέχει τα στοιχεία στην ετικέτα <item>. Τα πιο σημαντικά υποστοιχεία ενός αντικειμένου είναι:

1. ο τίτλος, ο οποίος περιέχει την επικεφαλίδα της ιστορίας
2. ο σύνδεσμος, ο οποίος είναι η διεύθυνση URL όπου μπορεί να ανακτηθεί ολόκληρη η ιστορία
3. η ημερομηνία και ώρα δημοσίευσης, pubDate, είναι η ημερομηνία και η ώρα που δημοσιεύεται η είδηση.
4. η κατηγορία, η οποία υποδηλώνει το είδος της είδησης ανάλογα την εκάστοτε ειδησεογραφική πύλη
5. η περιγραφή, η οποία περιέχει μερικές γραμμές σχετικά με την ιστορία και πολλές φορές είναι οι πρώτες προτάσεις της ιστορίας
6. το περιεχόμενο, το οποίο αναφέρει επιπρόσθετες πληροφορίες για την είδηση. Αυτό το υποστοιχείο μπορεί πολλές φορές να παραλείπεται.

3.2.4 Ιστορική ανασκόπηση – Σχετικές εργασίες

Υπάρχει πληθώρα ερευνητικών εργασιών που σχετίζονται και μελετούν το θέμα της απόξεσης ιστού. Οι εργασίες αυτές είτε συζητούν τεχνολογίες και εργαλεία που μπορούν να χρησιμοποιηθούν για την απόξεση ιστού σε ένα συγκεκριμένο πλαίσιο ή κλάδο, είτε σκιαγραφούν πιθανές ευκαιρίες που μπορεί να προσφέρει η απόξεση ιστού σε ερευνητές σε διάφορους κλάδους. Για παράδειγμα, η ερευνητική εργασία των [KT18] εξετάζει πώς η γλώσσα R, μαζί με τις βιβλιοθήκες της, μπορεί να χρησιμοποιηθεί για τη συλλογή, την οργάνωση και την προεπεξεργασία οικονομικών δεδομένων από τον Ιστό. Η εργασία των [BW17] δείχνει πώς τα δεδομένα που λαμβάνονται από το Craigslist μπορούν να ανακτηθούν και να αναλυθούν για να αποκτηθεί καλύτερη κατανόηση του τομέα ενοικίασης των ακινήτων στις Ηνωμένες Πολιτείες. Η εργασία του [K15] εξετάζει συγκεκριμένους τρόπους με τους οποίους μπορεί να χρησιμοποιηθεί η απόξεση ιστού, η εξόρυξη δεδομένων και οι οπτικοποιήσεις δεδομένων στον τομέα της δημοσιογραφίας. Συλλογικά, αυτές οι

εργασίες δείχνουν ότι υπάρχει αυξανόμενο ενδιαφέρον για η απόξεση ιστού σε πολλούς επιστημονικούς κλάδους και πως είναι ένας σημαντικός και αναπτυσσόμενος τομέας στον τομέα της Επιστήμης Υπολογιστών [KJS20].

3.3 Μέθοδοι καθαρισμού περιεχομένου άρθρων

Αφού εξαχθούν τα απαραίτητα δεδομένα από τις επιλεγμένες ειδησεογραφικές ιστοσελίδες, πρέπει να καθαριστούν, να υποβληθούν δηλαδή σε προεπεξεργασία και να οργανωθούν με τρόπο που να επιτρέπει την περαιτέρω ανάλυση τους. Η καλή προετοιμασία δεδομένων οδηγεί σε αποτελεσματική ανάλυση δεδομένων, εξαλείφοντας τα λάθη και ανακριβή αποτελέσματα κατά την επεξεργασία [DS22].

Το διαδικτυακό κείμενο μπορεί να περιέχει ανεπιθύμητο κείμενο, ακατέργαστα στοιχεία και άλλα αδιάφορα δεδομένα πληροφορίες, όπως διαφημίσεις [VRJD17]. Σύμφωνα με τον όγκο των δεδομένων που εμπλέκονται, είναι απαραίτητη μια προγραμματική προσέγγιση για εξοικονόμηση χρόνου. Υπάρχουν πολλές γλώσσες προγραμματισμού, όπως η R και η Python, οι οποίες περιέχουν βιβλιοθήκες επεξεργασίας φυσικής γλώσσας (Natural Language Processing (NLP)) για την επεξεργασία της φυσικής γλώσσας και λειτουργίες χειρισμού δεδομένων που είναι χρήσιμες για τον καθαρισμό και την οργάνωση των δεδομένων.

Η προεπεξεργασία των δεδομένων αποτελείται από τα παρακάτω βήματα: τμηματοποίηση, αφαίρεση λέξεων τερματισμού, αφαίρεση θορύβου και stemming ή λημματοποίηση, με αυτή τη σειρά.

3.3.1 Τμηματοποίηση

Η τμηματοποίηση (segmentation ή tokenization) είναι ένα θεμελιώδες βήμα προεπεξεργασίας στην επεξεργασία φυσικής γλώσσας, το οποίο περιλαμβάνει τη διάσπαση ενός τμήματος κειμένου σε μικρότερες μονάδες ή tokens. Αυτή η διαδικασία είναι σημαντική γιατί επιτρέπει την ανάλυση δεδομένων κειμένου με πιο δομημένο και οργανωμένο τρόπο. Χωρίς τμηματοποίηση, το κείμενο της φυσικής γλώσσας θα αντιμετωπίζεται ως μια μεγάλη ακολουθία χαρακτήρων, γεγονός που καθιστά δύσκολο για υπολογιστικά προγράμματα να εξάγουν σημαντικές πληροφορίες από αυτό.

Υπάρχουν διαφορετικές προσεγγίσεις για την τμηματοποίηση, καθεμία με τα δικά της πλεονεκτήματα και μειονεκτήματα.

- *Word-based tokenization*: βασίζεται σε λέξεις και είναι η πιο κοινή μορφή, στην οποία ένα κομμάτι κειμένου χωρίζεται σε μεμονωμένες λέξεις με βάση τα κενά και τα σημεία στίξης. Αυτή η προσέγγιση είναι απλή και αποτελεσματική για τις περισσότερες εργασίες NLP, αλλά μπορεί να έχει προκλήσεις για γλώσσες με πολύπλοκες μορφολογίες ή για δεδομένα κειμένου που περιέχουν μη τυπική ορθογραφία, όπως αναρτήσεις μέσων κοινωνικής δικτύωσης.
- *Character-based tokenization*: είναι μια εναλλακτική προσέγγιση, στην οποία ένα κομμάτι κειμένου χωρίζεται σε μεμονωμένους χαρακτήρες. Αυτή η προσέγγιση μπορεί να είναι χρήσιμη για γλώσσες με μη αλφαβητικά δεδομένα, αλλά μπορεί να οδηγήσει σε αυξημένη υπολογιστική πολυπλοκότητα.
- *Subword-based tokenization*: είναι μια άλλη προσέγγιση, στην οποία ένα κομμάτι κειμένου χωρίζεται σε μικρότερες ενότητες, όπως προθέματα, επιθήματα και λέξεις ρίζας. Αυτή η προσέγγιση μπορεί να είναι αποτελεσματική για γλώσσες με πολύπλοκα συστήματα κλίσης, καθώς επιτρέπει στις μηχανές να καταγράφουν την εσωτερική δομή των λέξεων. Ωστόσο, το tokenization που βασίζεται σε υπολέξεις μπορεί να είναι υπολογιστικά ακριβό σε σχέση με το tokenization που βασίζεται σε λέξεις.

Η επιλογή της προσέγγισης του tokenization εξαρτάται από τη συγκεκριμένη εργασία και τα χαρακτηριστικά των δεδομένων κειμένου. Το πιο συχνά χρησιμοποιούμενο εργαλείο για τη τμηματοποίηση, είναι το Natural Language Toolkit (NLTK). [AZHL22].

3.3.2 Αφαίρεση θορύβου

Η αφαίρεση θορύβου στην προεπεξεργασία δεδομένων κειμένου αναφέρεται στην αφαίρεση ειδικών συμβόλων ή ψηφίων που δεν σχετίζονται με την ανάλυση ή την κατηγοριοποίηση του κειμένου, όπως [,०-৯,0-9,a-z,@,!,#,$,%,&,.,„,|,:,],+ κ.λπ. Αυτό γίνεται για να μειωθεί η ποσότητα του θορύβου ή οι άσχετες πληροφορίες που ενδέχεται να επηρεάσουν την ακρίβεια της ανάλυσης.

Ορισμένα κοινά παραδείγματα συμβόλων ή χαρακτήρων που απομακρύνονται κατά την αφαίρεση θορύβου περιλαμβάνουν ειδικούς χαρακτήρες όπως αγκύλες, σημεία στίξης, ψηφία και άλλους μη αλφαβητικούς χαρακτήρες. Αφαιρώντας αυτά τα σύμβολα, το κείμενο μπορεί να τυποποιηθεί και να γίνει πιο συνεπές, καθιστώντας ευκολότερη την επεξεργασία και την ανάλυση του.

Ο πρωταρχικός στόχος της αφαίρεσης θορύβου είναι η τυποποίηση του κειμένου και η συνοχή του σε όλο το σύνολο δεδομένων. Αυτό, με τη σειρά του, βοηθά στη βελτίωση της ακρίβειας και της αξιοπιστίας της ανάλυσης που εκτελείται στα δεδομένα κειμένου. Η αφαίρεση θορύβου είναι ιδιαίτερα σημαντική στην ταξινόμηση κειμένου και σε άλλες εργασίες επεξεργασίας φυσικής γλώσσας [BSH21].

3.3.3 Εξάλειψη των λέξεων τερματισμού (stop words)

Η αφαίρεση των stop words είναι ένα κρίσιμο βήμα προεπεξεργασίας στην επεξεργασία φυσικής γλώσσας, το οποίο περιλαμβάνει την αφαίρεση λέξεων που θεωρούνται κοινές ή ασήμαντες μέσα σε ένα κείμενο. Ο στόχος είναι να μειωθεί το μέγεθος των δεδομένων και να βελτιωθεί η αποτελεσματικότητα και η ακρίβεια της ανάλυσης που ακολουθεί. Τα stop words ορίζονται συνήθως ως λέξεις που εμφανίζονται συχνά σε μια γλώσσα, όπως "το", "και", "ένα" και "σε". Αυτές οι λέξεις εξυπηρετούν γραμματικούς σκοπούς σε μια πρόταση, αλλά δεν έχουν μεγάλη σημασιολογική συσχέτιση.

Υπάρχουν διαφορετικές προσεγγίσεις για την αφαίρεση των stop words, ανάλογα με την απαιτούμενη εργασία και τα χαρακτηριστικά του κειμένου που αναλύεται. Μια κοινή προσέγγιση είναι η χρήση μιας προκαθορισμένης λίστας stop words, η οποία συχνά βασίζεται σε ένα σύνολο δεδομένων κειμένου για μια συγκεκριμένη γλώσσα ή τομέα. Μια άλλη προσέγγιση είναι η χρήση στατιστικών μεθόδων για τον εντοπισμό λέξεων που εμφανίζονται συχνά σε ένα σύνολο δεδομένων και την αφαίρεση τους, καθώς θεωρούνται stop words [DZ11].

3.3.4 Μετατροπή σε βασική μορφή λέξεων (Stemming)

Το stemming είναι μια τεχνική που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας για τη τροποποίηση των λέξεων στη ριζική τους μορφή αφαιρώντας προθέματα και επιθήματα. Η ριζική μορφή που προκύπτει μπορεί να μην είναι απαραίτητα μια πραγματική λέξη, αλλά εξακολουθεί να αντιπροσωπεύει το βασικό νόημα της αρχικής λέξης.

Ο πρωταρχικός στόχος του stemming είναι να απλοποιήσει το κείμενο και να μειώσει τον αριθμό των μοναδικών λέξεων που πρέπει να υποστούν επεξεργασία. Αυτή η απλοποίηση μπορεί να βελτιώσει την αποτελεσματικότητα των αλγορίθμων και των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για εργασίες επεξεργασίας φυσικής γλώσσας, όπως η ταξινόμηση κειμένου.

Οι βασικοί αλγόριθμοι λειτουργούν εφαρμόζοντας ένα σύνολο προκαθορισμένων κανόνων ή αλγορίθμων στο κείμενο για τον εντοπισμό και την αφαίρεση επιθεμάτων. Οι κανόνες μπορεί να διαφέρουν ανάλογα με τη γλώσσα που υποβάλλεται σε επεξεργασία και διαφορετικοί αλγόριθμοι μπορεί να παράγουν διαφορετικά αποτελέσματα.

Εικόνα 9: Τμηματοποίηση. Πηγή: turing.com

Για το stemming υπάρχουν πολλοί stemmers [KB16] όπως:

- *S-Stemmer*: σχετίζεται με τις λέξεις των οποίων το μήκος είναι μεγαλύτερο από τρία γράμματα. Το κίνητρο αυτού του stemmer είναι να λαμβάνει υπόψη και τον ενικό και τον πληθυντικό αριθμό.
- *Lovins Stemmer*: ήταν ο πρώτος stemmer που προτάθηκε. Το μεγαλύτερό του πλεονέκτημα είναι η ταχύτητά του. Αυτός ο stemmer έχει 294 καταλήξεις, 29 συνθήκες και 35 κανόνες μετασχηματισμού.
- *Porter Stemmer(?)*: αυτός είναι ο πιο ευρέως χρησιμοποιούμενος stemmer λόγω της υψηλής ακρίβειας και του απλού αλγορίθμου του. Περιλαμβάνει πέντε εύκολα βήματα τα οποία εφαρμόζονται σε κάθε λέξη.
- *Paice/Husk stemmer*: αυτός ο stemmer υλοποιεί έναν αλγόριθμο που βασίζεται στην επανάληψη και υλοποιεί τους ίδιους κανόνες και επιθήματα για κάθε βρόχο. Κάθε κανόνας έχει πέντε βήματα, μεταξύ των οποίων τρία είναι υποχρεωτικά να εφαρμοστούν ενώ τα υπόλοιπα δύο είναι προαιρετικά.

Αν και το stemming μπορεί να είναι μια αποτελεσματική τεχνική για την απλοποίηση του κειμένου, μπορεί επίσης να εισάγει σφάλματα ή ανακρίβειες στο κείμενο. Για παράδειγμα, το stemming μπορεί να παράγει βασικές μορφές που στην πραγματικότητα δεν υπάρχουν στη γλώσσα ή μπορεί να αποτύχει να εντοπίσει λεπτές παραλλαγές στη σημασία των λέξεων. Ως αποτέλεσμα, το stemming χρησιμοποιείται

συχνά σε συνδυασμό με άλλες τεχνικές, όπως η λημματοποίηση, για τη βελτίωση της ακρίβειας και της αξιοπιστίας των εργασιών επεξεργασίας φυσικής γλώσσας.

3.3.5 Λημματοποίηση

Το τελευταίο και πιο κρίσιμο βήμα για την προεπεξεργασία του κειμένου είναι η λημματοποίηση (lemmatization). Όπως τα παραπάνω, είναι ένα βήμα προεπεξεργασίας στην επεξεργασία φυσικής γλώσσας και η διαδικασία της περιλαμβάνει την αναγωγή των λέξεων στη βασική τους ρίζα αφαιρώντας τα συνημμένα επιθέματα των λέξεων [AZHL22]. Η λημματοποίηση είναι παρόμοια με το stemming, το οποίο περιλαμβάνει επίσης τη μετατροπή των λέξεων στη βασική τους μορφή. Ωστόσο, η λημματοποίηση λαμβάνει υπόψη το νόημα της πρότασης και το μέρος του λόγου μιας λέξης, ενώ το stemming απλώς αφαιρεί τα επιθήματα μιας λέξης για να αποκτήσει τη ριζική της μορφή [DZ11].

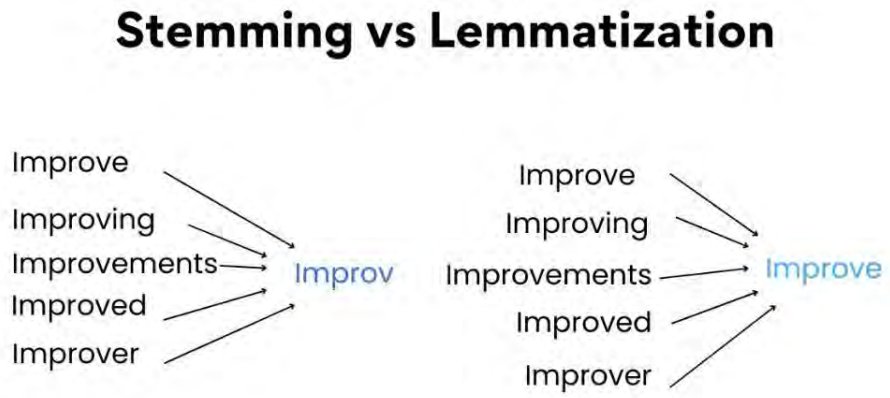
Εικόνα 10: Λημματοποίηση. Πηγή: turing.com

Η λημματοποίηση μπορεί να πραγματοποιηθεί χρησιμοποιώντας μια προσέγγιση που βασίζεται είτε σε κανόνες είτε σε λεξικό.

- Στη *λημματοποίηση βάσει κανόνων (rule-based)*, ένα σύνολο κανόνων εφαρμόζεται στα δεδομένα κειμένου για να προσδιορίσει τη βασική μορφή μιας λέξης με βάση το μέρος του λόγου στην οποία ανήκει.
- Στη *λημματοποίηση βάσει λεξικού (dictionary-based),* χρησιμοποιείται ένα προκατασκευασμένο λεξικό λημμάτων και των σχετικών μορφών για την εκτέλεση της λημματοποίησης. Αυτή η προσέγγιση μπορεί να είναι πιο ακριβής από τη λημματοποίηση που βασίζεται σε κανόνες, αλλά μπορεί να απαιτεί μεγαλύτερο λεξικό και περισσότερους υπολογιστικούς πόρους.

Γενικά, η λημματοποίηση παράγει πιο ακριβή και ουσιαστικά αποτελέσματα από το stemming, επειδή λαμβάνει υπόψη το πλαίσιο της λέξης και το μέρος του λόγου της. Ωστόσο, η λημματοποίηση είναι πιο εντατική και χρονοβόρα από υπολογισμούς

του stemming, το οποίο είναι μια ταχύτερη και απλούστερη προσέγγιση που χρησιμοποιείται συχνά σε εφαρμογές όπου η ταχύτητα και η απλότητα είναι πιο σημαντικές από την ακρίβεια.



Εικόνα 11: Σύγκριση μετατροπής στη βασική μορφή λέξεων και της λημματοποίησης. Πηγή: turing.com

3.4 Επίλογος

Ο Παγκόσμιος Ιστός παρέχει μεγάλο όγκο πληροφοριών που αυξάνεται συνεχώς, και η πρόσβαση σε αυτές τις πληροφορίες αποτελεί σημαντικό μέρος της καθημερινής ζωής μας. Οι ειδήσεις αποτελούν σημαντικό κομμάτι των δεδομένων που διατίθενται στον Παγκόσμιο Ιστό, και η διαχείριση τους και η εξαγωγή χρήσιμων πληροφοριών απαιτούν αυτοματοποιημένες τεχνικές, όπως η απόξεση ιστού. Το βήμα αυτό όμως δεν είναι αρκετό, καθώς τα δεδομένα που εξάγονται από την απόξεση ιστού περιέχουν πληροφορίες που δεν έχουν κάποια σημασία, όπως οι διαφημίσεις αλλά και λέξεις χωρίς κάποια ουσιαστική σημασία. Επομένως, κρίνεται απαραίτητη η προεπεξεργασία των ακατέργαστων πληροφοριών προκειμένου να είναι εφικτή η ανάλυση μόνο των χρήσιμων πληροφοριών μειώνοντας τον χρόνο επεξεργασίας των δεδομένων και αυξάνοντας την αποδοτικότητα του συστήματος. Στην καλή προετοιμασία δεδομένων συμπεριλαμβάνεται κατά σειρά η τμηματοποίηση, η αφαίρεση λέξεων τερματισμού, η αφαίρεση θορύβου, το stemming ή/και η λημματοποίηση. Με τον τρόπο αυτό, γίνεται περαιτέρω αποτελεσματική ανάλυση των δεδομένων, απομακρύνοντας σφάλματα και ανακρίβειες που μπορεί να προκύψουν κατά την επεξεργασία.

4 Μέθοδοι κατάταξης (classification) ειδησεογραφικών άρθρων βάσει περιεχομένου

4.1 Εισαγωγή

Οι άνθρωποι βασίζονται στις ειδήσεις για να γνωρίζουν τι συμβαίνει γύρω τους αλλά και στον κόσμο, προκειμένου να ενημερώνονται την καθημερινότητά τους. Στον σημερινό κόσμο, όπου η διάδοση των ψευδών ειδήσεων είναι ανεξέλεγκτη, μια μεγάλης κλίμακας και υψηλής ποιότητας πηγή αυθεντικών ειδήσεων είναι πολύτιμη.

Μετά τη προεπεξεργασία των ειδήσεων, ακολουθεί η ταξινόμηση των ειδήσεων στις κατάλληλες κατηγορίες [M22]. Η αυτόματη ταξινόμηση κειμένων έχει κερδίσει την προσοχή τα τελευταία χρόνια λόγω της πληθώρας αδόμητων δεδομένων κειμένου, ιδιαίτερα στο πλαίσιο του περιεχομένου ειδήσεων στο διαδίκτυο. Ωστόσο, η διαδικασία αυτή βρίσκεται στα αρχικά της στάδια [BSH21].

Η ταξινόμηση κειμένου είναι μια θεμελιώδης εργασία στην επεξεργασία φυσικής γλώσσας (NLP) που περιλαμβάνει την αυτόματη αντιστοίχιση προκαθορισμένων κατηγοριών με βάση το περιεχόμενο κειμένου και τα εξαγόμενα χαρακτηριστικά του [DZ11]. Διαδραματίζει κρίσιμο ρόλο σε διάφορες εφαρμογές, όπως της ανάκτησης πληροφοριών, της ανάλυσης συναισθημάτων, των συστημάτων συστάσεων και άλλων. Με την κατηγοριοποίηση δεδομένων κειμένου, η ταξινόμηση τους επιτρέπει την αποτελεσματική οργάνωση, ανάκτηση και ανάλυση πληροφοριών, οδηγώντας σε πολλά σημαντικά οφέλη.

Συγκεκριμένα, η κατηγοριοποίηση των ειδήσεων είναι μια σημαντική εφαρμογή της ταξινόμησης κειμένων, η οποία περιλαμβάνει την κατανομή των άρθρων ειδήσεων σε προκαθορισμένες κατηγορίες με βάση το περιεχόμενό τους.

Η κατηγοριοποίηση ειδήσεων επιτρέπει την αποτελεσματική οργάνωση μεγάλου όγκου άρθρων ειδήσεων, διευκολύνοντας τους χρήστες να περιηγηθούν και να έχουν πρόσβαση στον συγκεκριμένο τύπο πληροφοριών που τους ενδιαφέρουν. Κατηγορίες όπως πολιτική, αθλητικά, τεχνολογία, ψυχαγωγία και επιχειρήσεις παρέχουν έναν δομημένο τρόπο πλοήγησης στο ποικίλο τοπίο ειδήσεων.

Επιπλέον, καθίσταται δυνατή η εξατομικευμένη παράδοση ειδήσεων. Οι χρήστες μπορούν να καθορίσουν τις προτιμήσεις τους και να λαμβάνουν άρθρα

ειδήσεων προσαρμοσμένα στα ενδιαφέροντά τους. Αυτό βελτιώνει την εμπειρία του χρήστη παρέχοντας σχετικό και στοχευμένο περιεχόμενο.

Ακόμα, η κατηγοριοποίηση ειδήσεων επιτρέπει την ανάπτυξη συστημάτων προτάσεων που προτείνουν σχετικά άρθρα ειδήσεων στους χρήστες με βάση το ιστορικό ανάγνωσης και τις προτιμήσεις τους. Βοηθά στο φιλτράρισμα άσχετων ή ανεπιθύμητων άρθρων ειδήσεων, μειώνοντας την υπερφόρτωση πληροφοριών και βελτιώνοντας την ποιότητα της εμπειρίας κατανάλωσης ειδήσεων.

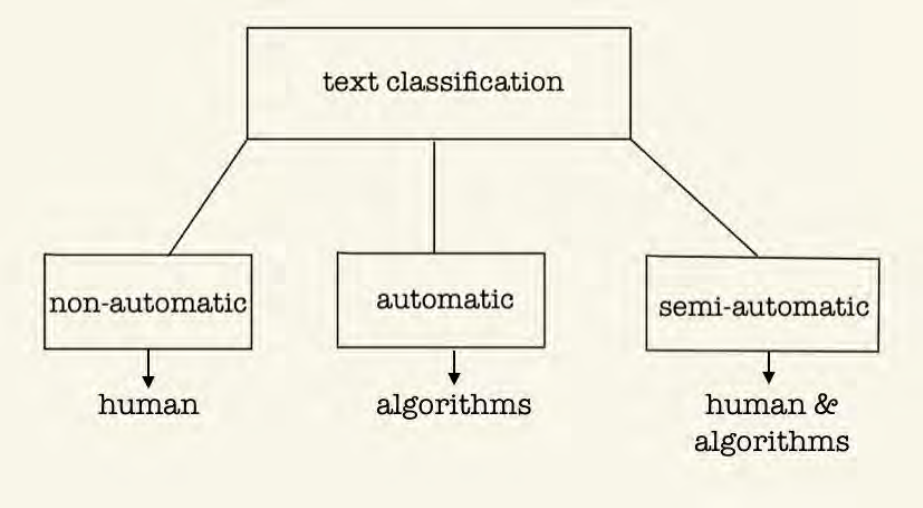
Οργανισμοί, πρακτορεία μέσων ενημέρωσης και ερευνητές συχνά χρειάζεται να παρακολουθούν το περιεχόμενο ειδήσεων για διάφορους σκοπούς. Μερικοί από αυτούς περιλαμβάνουν την παρακολούθηση του κοινού αισθήματος τη διεξαγωγή ανταγωνιστικών αναλύσεων. Η κατηγοριοποίηση των ειδήσεων επιτρέπει την αποτελεσματική παρακολούθηση ταξινομώντας αυτόματα τα εισερχόμενα άρθρα ειδήσεων σε σχετικές κατηγορίες, διευκολύνοντας την ανάλυση δεδομένων και τις διαδικασίες λήψης αποφάσεων.

Η ανάλυση της διανομής και των τάσεων των άρθρων ειδήσεων σε διαφορετικές κατηγορίες μπορεί να προσφέρει πολύτιμες πληροφορίες για την κοινωνική, πολιτιστική και οικονομική δυναμική. Επιτρέπει σε ερευνητές και αναλυτές να εντοπίζουν τις αναδυόμενες τάσεις, να παρακολουθούν την κοινή γνώμη και να κατανοούν τον αντίκτυπο συγκεκριμένων γεγονότων ή θεμάτων.

Η κατηγοριοποίηση ειδήσεων ωφελεί επίσης τους διαφημιστές, παρέχοντας πληροφορίες για το πλαίσιο και το κοινό-στόχο των άρθρων ειδήσεων. Οι διαφημιστές μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για να ευθυγραμμίσουν καλύτερα τις διαφημιστικές τους καμπάνιες με συγκεκριμένες κατηγορίες και να στοχεύσουν σχετικό κοινό, μεγιστοποιώντας την αποτελεσματικότητα των τοποθετήσεων διαφημίσεών τους.

4.2 Μέθοδοι ταξινόμησης ειδήσεων

Οι ερευνητές έχουν αναπτύξει διάφορες τεχνικές για την ταξινόμηση των ειδήσεων, αλλά η επιδίωξη για μεγαλύτερη ακρίβεια σε αυτές τις τεχνικές παραμένει συνεχής [KM23]. Διακρίνουμε τρεις κατηγορίες ταξινόμησης: τη μη αυτόματη ή χειροκίνητη, την ημι-αυτόματη και την αυτόματη.



Εικόνα 12: Κατηγορίες Ταξινόμησης

4.2.1 Μη αυτόματη ταξινόμηση

Η μη αυτόματη (χειροκίνητη) ταξινόμηση γεγονότων από ρεπορτάζ ειδήσεων αναφέρεται στη διαδικασία ταξινόμησης ή οργάνωσης δεδομένων κειμένου με μη αυτόματο τρόπο [CW15]. Δηλαδή, χωρίς τη βοήθεια αυτοματοποιημένων εργαλείων ή αλγορίθμων γεγονός που τη καθιστά εξαιρετικά δαπανηρή και χρονοβόρα. Περιλαμβάνει ανθρώπινη παρέμβαση και τεχνογνωσία για την ανάλυση και την ανάθεση κατάλληλων κατηγοριών ή ετικετών στο κείμενο με βάση το περιεχόμενο, το πλαίσιο και τον σκοπό του.

Η ταξινόμηση κειμένου συχνά περιλαμβάνει χειροκίνητο σχολιασμό, όπου οι ταξινομητές διαβάζουν και αναλύουν τα έγγραφα κειμένου και τους αναθέτουν κατάλληλες κατηγορίες ή ετικέτες. Αυτή η διαδικασία απαιτεί εξειδίκευση και εξοικείωση με το αντικείμενο ή το σχήμα ταξινόμησης. Οι άνθρωποι αυτοί, ερμηνεύουν το νόημα και το πλαίσιο του κειμένου για να λάβουν ενημερωμένες αποφάσεις ταξινόμησης, γεγονός όμως που το καθιστά υποκειμενικό για τον κάθε ταξινομητή. Διαφορετικοί σχολιαστές μπορεί να ερμηνεύουν το ίδιο κείμενο διαφορετικά, οδηγώντας σε παραλλαγές στην ταξινόμηση.

4.2.2 Αυτόματη ταξινόμηση

Η αυτοματοποιημένη ταξινόμηση κειμένου αναφέρεται στη διαδικασία κατηγοριοποίησης ή οργάνωσης δεδομένων κειμένου χρησιμοποιώντας υπολογιστικές τεχνικές και αλγόριθμους χωρίς σημαντική ανθρώπινη παρέμβαση. Περιλαμβάνει τη χρήση μηχανικής μάθησης, επεξεργασίας φυσικής γλώσσας (NLP) και άλλων

αυτοματοποιημένων μεθόδων για την ανάλυση και την ταξινόμηση κειμένου με βάση το περιεχόμενό του, το πλαίσιο ή άλλα σχετικά χαρακτηριστικά.

Η αυτοματοποιημένη ταξινόμηση κειμένου συχνά χρησιμοποιεί εποπτευόμενους αλγόριθμους μηχανικής μάθησης. Αυτοί οι αλγόριθμοι εκπαιδεύονται σε ένα σύνολο δεδομένων, όπου κάθε έγγραφο κειμένου σχετίζεται με μια γνωστή κατηγορία ή ετικέτα. Τα μοντέλα μαθαίνουν μοτίβα και σχέσεις στα δεδομένα και χρησιμοποιούν αυτή τη γνώση για να ταξινομήσουν νέα, άγνωστα κείμενα. Υπάρχουν διάφοροι αλγόριθμοι μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν, όπως Naive Bayes, Support Vector Machines (SVM), δέντρα αποφάσεων, random forest και νευρωνικών δικτύων. Η επιλογή του αλγορίθμου εξαρτάται από παράγοντες όπως η φύση των δεδομένων, το μέγεθος του συνόλου δεδομένων και οι συγκεκριμένες απαιτήσεις της εργασίας ταξινόμησης.

Ενώ η εποπτευόμενη μάθηση χρησιμοποιείται συνήθως, μέθοδοι χωρίς επίβλεψη και ημι-επίβλεψη μπορούν επίσης να χρησιμοποιηθούν. Οι μη εποπτευόμενες τεχνικές στοχεύουν στην ανακάλυψη μοτίβων και συστάδων στα δεδομένα κειμένου χωρίς προκαθορισμένες ετικέτες-κατηγορίες, ενώ οι ημι-εποπτευόμενες μέθοδοι συνδυάζουν δεδομένα με ετικέτα και χωρίς ετικέτα για να βελτιώσουν την απόδοση ταξινόμησης.

Επιπλέον, η αυτόματη ταξινόμηση μπορεί να αξιοποιεί την εξαγωγή σχετικών χαρακτηριστικών από τα δεδομένα κειμένου για την αναπαράσταση του περιεχομένου του. Αυτό μπορεί να περιλαμβάνει τεχνικές όπως Bag-of-Words, n-grams, word embeddings ή άλλες στατιστικές αναπαραστάσεις. Αυτά τα χαρακτηριστικά καταγράφουν τα βασικά χαρακτηριστικά του κειμένου, επιτρέποντας στον αλγόριθμο να μάθει μοτίβα για ταξινόμηση.

Η αυτοματοποιημένη ταξινόμηση κειμένου μπορεί επίσης να αξιοποιήσει προεκπαιδευμένα μοντέλα σε προσεγγίσεις transfer learning. Αυτά τα μοντέλα εκπαιδεύονται σε σύνολα δεδομένων μεγάλης κλίμακας για γενική κατανόηση της γλώσσας και μπορούν να βελτιστοποιηθούν ή να προσαρμοστούν για συγκεκριμένες εργασίες ταξινόμησης. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη όταν υπάρχουν περιορισμένα διαθέσιμα δεδομένα με γνωστή ετικέτα.

Παρόλα αυτά, η συγκεκριμένη κατηγορία ταξινόμησης έχει περιορισμούς όσον αφορά την ακρίβεια και την ευαισθησία των δεδομένων που συλλέγονται. Η χρήση

αυτοματοποιημένων μεθόδων για την ανάλυση κειμένου απαιτεί από τον αναλυτή να λάβει πολλαπλές αποφάσεις που συχνά λαμβάνονται ελάχιστα αλλά έχουν συνέπειες που δεν είναι ούτε προφανείς, ούτε καλοήθεις [CW15].

4.2.3 Ημι-αυτόματη ταξινόμηση

Η ημιαυτόματη ταξινόμηση κειμένου αναφέρεται σε μια υβριδική προσέγγιση που συνδυάζει τα πλεονεκτήματα τόσο των αυτοματοποιημένων αλγορίθμων όσο και της ανθρώπινης τεχνογνωσίας στη διαδικασία κατηγοριοποίησης ή οργάνωσης δεδομένων κειμένου. Περιλαμβάνει τη χρήση αυτοματοποιημένων τεχνικών για να βοηθήσει τους ανθρώπους στη διαδικασία ταξινόμησης, βελτιώνοντας την αποτελεσματικότητα και την ακρίβεια, αντιμετωπίζοντας παράλληλα, τις προκλήσεις της χειροκίνητης και της πλήρως αυτόματης ταξινόμησης.

Ο κύριος στόχος της ημιαυτόματης διαδικασίας ταξινόμησης είναι να εξαλείψει τόσα πολλά από αυτά τα «άσχετα» άρθρα ειδήσεων, διατηρώντας όσο το δυνατόν περισσότερα από όλα τα «σχετικά» άρθρα έτσι ώστε να δοθούν σε εκπαιδευμένο άνθρωπο για περαιτέρω ταξινόμηση [CW15]. Μετά την αρχική ταξινόμηση από τον επιλεγμένο αλγόριθμο πχ. μηχανική μάθηση, οι ειδικοί στον άνθρωπο εξετάζουν και επικυρώνουν τα αποτελέσματα. Εξετάζουν τις ταξινομήσεις που γίνονται από τον αλγόριθμο και κάνουν τις απαραίτητες διορθώσεις ή προσαρμογές με βάση την τεχνογνωσία και τις γνώσεις τους στον τομέα. Αυτή η επαναληπτική διαδικασία βοηθά στη βελτίωση των αποτελεσμάτων ταξινόμησης και στη βελτίωση της ακρίβειας.

Η ημιαυτόματη ταξινόμηση κειμένου προσφέρει μια ισορροπία μεταξύ αποτελεσματικότητας και ακρίβειας. Αυτή η προσέγγιση μειώνει τον φόρτο εργασίας χειροκίνητης ταξινόμησης, διατηρώντας παράλληλα υψηλό επίπεδο συνάφειας στα κωδικοποιημένα δεδομένα [CW15]. Είναι ιδιαίτερα χρήσιμη όταν ασχολούμαστε με μεγάλους όγκους δεδομένων κειμένου, όπου η πλήρως χειροκίνητη ταξινόμηση μπορεί να είναι μη πρακτική ή χρονοβόρα.

4.2.3.1 Ταξινόμηση με χρήση keywords

Η ταξινόμηση κειμένων με χρήση λέξεων-κλειδιών είναι μια κοινώς χρησιμοποιούμενη προσέγγιση όπου τα έγγραφα κατηγοριοποιούνται με βάση την παρουσία ή την απουσία συγκεκριμένων λέξεων-κλειδιών ή φράσεων-κλειδιών. Περιλαμβάνει τη δημιουργία ενός συνόλου keywords που είναι ενδεικτικές διαφορετικών κατηγοριών και στη συνέχεια την αντιστοίχιση αυτών των λέξεων-

κλειδιών με το κείμενο για να προσδιοριστεί η ταξινόμησή του. Οι πληροφορίες σχετικά με τα μεταδεδομένα είναι ιδιαίτερα χρήσιμες στην ταξινόμηση όπως, ειδικά ουσιαστικά όπως ονόματα προσώπων/τόπων, τίτλος εγγράφου, όνομα συγγραφέα του εγγράφου κ.λπ. [DZ11].

Αρχικά ορίζεται ένα σύνολο λέξεων-κλειδιών ή φράσεων-κλειδιών που σχετίζονται με την εργασία ταξινόμησης και πρέπει να αποτυπώνουν τα σημαντικά θέματα ή χαρακτηριστικά κάθε κατηγορίας. Μπορεί να είναι όροι διαφόρων τομέων, συγκεκριμένες λέξεις που σχετίζονται με ορισμένα θέματα ή φράσεις που αντικατοπτρίζουν συναίσθημα ή πρόθεση. Έπειτα, τα έγγραφα κειμένου υποβάλλονται σε επεξεργασία και ελέγχεται η παρουσία ή η απουσία λέξεων-κλειδιών. Αυτό μπορεί να γίνει εκτελώντας μια απλή αναζήτηση συμβολοσειρών, όπου το κείμενο σαρώνεται για ακριβείς αντιστοιχίσεις των λέξεων-κλειδιών. Εναλλακτικά, μπορούν να χρησιμοποιηθούν περίπλοκες τεχνικές όπως οι κανονικές εκφράσεις ή η επεξεργασία φυσικής γλώσσας (NLP) για τον χειρισμό παραλλαγών σε μορφές λέξεων, συνώνυμα ή σχετικούς όρους.

Εάν ένα έγγραφο περιέχει επαρκή αριθμό λέξεων-κλειδιών από μια συγκεκριμένη κατηγορία, εκχωρείται σε αυτήν την κατηγορία. Ο αριθμός των λέξεων-κλειδιών που απαιτούνται για την ταξινόμηση μπορεί να προκαθοριστεί ή να προσδιοριστεί εμπειρικά με βάση το επιθυμητό επίπεδο ακρίβειας και ανάκλησης. Η ταξινόμηση κειμένων με χρήση λέξεων-κλειδιών μπορεί να εφαρμοστεί σε πολλές κατηγορίες ταυτόχρονα. Σε τέτοιες περιπτώσεις, κάθε κατηγορία έχει το σύνολο των λέξεων-κλειδιών της και το κείμενο αξιολογείται σε σχέση με όλες τις κατηγορίες. Το έγγραφο μπορεί να αντιστοιχιστεί σε πολλές κατηγορίες, εάν ταιριάζει με λέξεις-κλειδιά από πολλά σύνολα.

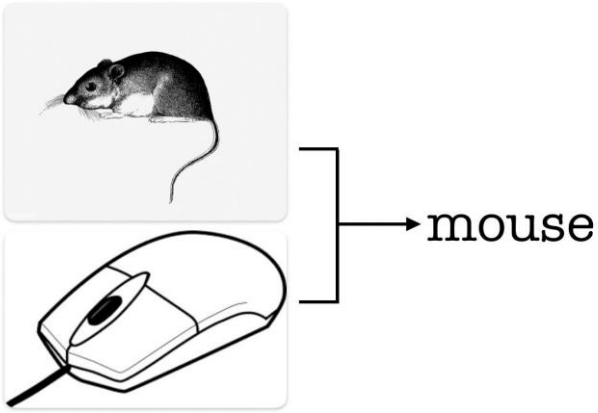
Παρόλο που η ταξινόμηση κειμένου με χρήση λέξεων-κλειδιών είναι απλή και ερμηνεύσιμη, έχει ορισμένους περιορισμούς. Βασίζεται σε μεγάλο βαθμό στην επιλογή των λέξεων-κλειδιών και εάν αυτές δεν είναι σωστά επιλεγμένες, μπορεί να οδηγήσει σε ανακριβείς ταξινομήσεις. Επιπλέον, αυτή η προσέγγιση μπορεί να δυσκολεύεται να χειριστεί παραλλαγές σε μορφές λέξεων, έννοιες που εξαρτώνται από το περιβάλλον ή καταστάσεις όπου το έγγραφο δεν περιέχει ρητές αντιστοιχίσεις λέξεων-κλειδιών.

Η ταξινόμηση κειμένων με χρήση λέξεων-κλειδιών μπορεί να συνδυαστεί με άλλες προσεγγίσεις για βελτιωμένη ακρίβεια. Για παράδειγμα, η ταξινόμηση βάσει λέξεων-κλειδιών μπορεί να χρησιμεύσει ως βήμα προεπεξεργασίας για το φιλτράρισμα εγγράφων που είναι πιθανό να ανήκουν σε μια συγκεκριμένη κατηγορία, ακολουθούμενη από πιο εξελιγμένες μεθόδους, για περαιτέρω βελτίωση της ταξινόμησης.

Δεδομένου του γεγονότος ότι οι λέξεις-κλειδιά που χρησιμοποιούνται είναι εξαιρετικά κοινές, η συντριπτική πλειοψηφία των ανακτημένων άρθρων δεν περιέχει σχετικές πληροφορίες. Έτσι, ο αριθμός των άσχετων άρθρων υπερβαίνει κατά πολύ εκείνα που καλύπτουν γεγονότα που παρουσιάζουν ενδιαφέρον για ανάλυση [CW15]. Η ύπαρξη περισσότερων άρθρων δεν σημαίνει απαραίτητα ότι το ένα σώμα είναι καλύτερο από το άλλο. Η έλλειψη επικάλυψης μπορεί να υποδεικνύει ότι η αναζήτηση της κατηγορίας θέματος είναι πολύ στενή ή/και η αναζήτηση λέξεων-κλειδιών είναι πολύ ευρεία. Στο σώμα των εγγράφων, χρειάζεται να περιλαμβάνονται όλα τα σχετικά αντικείμενα για να ελαχιστοποιηθούν τα ψευδώς αρνητικά αποτελέσματα και να αποκλειστούν τυχόν άσχετα αντικείμενα για να ελαχιστοποιηθούν τα ψευδώς θετικά.

Με κάθε απόφαση σχετικά με το ποιες λέξεις ή όρους να συμπεριλάβουμε ή να παραλείψουμε από μια αναζήτηση λέξεων-κλειδιών, διατρέχουμε τον κίνδυνο να εισαγάγουμε μεροληψία [B+21]. Ωστόσο, η προσεκτική εφαρμογή της επέκτασης λέξεων-κλειδιών μπορεί να ελαχιστοποιήσει την πιθανότητα αυτού του τύπου σφάλματος. Εν ολίγοις, ο αναλυτής θα πρέπει να προσπαθήσει για μια αναζήτηση λέξεων-κλειδιών που μεγιστοποιεί τόσο τη συνάφεια όσο και την αντιπροσώπευση έναντι του πληθυσμού που ενδιαφέρει.

Υπάρχουν πολλοί λόγοι για τους οποίους οποιαδήποτε αναζήτηση λέξης-κλειδιού μπορεί να είναι προβληματική: οι σχετικοί όροι μπορεί να αλλάξουν με την πάροδο του χρόνου, διαφορετικές δημοσιεύσεις μπορούν να χρησιμοποιούν συνώνυμα που παραβλέπονται και ούτω καθεξής. Γενικά, δύο βασικά προβλήματα συναντώνται στην ταξινόμηση, η «πολυσημία», δηλαδή μία λέξη έχει πολλές διαφορετικές σημασίες και η «συνωνυμία» όπου διαφορετικές λέξεις έχουν την ίδια σημασία πχ. γρήγορα, γοργά. Ωστόσο, οι αναζητήσεις λέξεων-κλειδιών είναι υπό τον έλεγχο του αναλυτή, είναι διαφανείς, αναπαραγώγιμες και φορητές.



Εικόνα 13: Παράδειγμα Πολυσημίας

4.3 Μέθοδοι κατάταξης ειδησεογραφικών άρθρων στα πληροφοριακά συστήματα καταγραφής γεγονότων

Ύστερα από βιβλιογραφική ανασκόπηση παρόμοιων εργασιών στην ταξινόμηση κειμένου, μπορούμε να διακρίνουμε δύο διαφορετικές προσεγγίσεις. Μία προσέγγιση περιλαμβάνει χειροκίνητη ανάγνωση και κωδικοποίηση συμβάντων από εκπαιδευμένους ανθρώπους που αποτελούν κωδικοποιητές, όπως για παράδειγμα από την Uppsala Conflict Data Program's Geo-referenced Event Dataset [SM13] ή Armed Conflict Location and Events Dataset [ACLED]. Η άλλη προσέγγιση, η αυτοματοποιημένη κωδικοποίηση, χρησιμοποιεί διάφορους αλγόριθμους για εξαγωγή συμβάντων υπολογιστικά από το κείμενο προέλευσης, όπως για το Kansas Event Data System [SH98] ή το πρόσφατο ICEWS event dataset [W+13] [CW15].

4.4 Επίλογος

Η κατηγοριοποίηση ειδήσεων αποτελεί ένα σημαντικό και αναπόσπαστο κομμάτι της επεξεργασίας φυσικής γλώσσας και έχει πολλαπλά οφέλη στη σύγχρονη κοινωνία μας. Με τη βοήθεια διαφόρων ειδών αλγορίθμων ταξινόμησης κειμένου, δημιουργείται μια πηγή αξιόπιστων ειδήσεων, η οποία είναι πολύτιμη για τους ανθρώπους που επιθυμούν να ενημερώνονται για τα τρέχοντα γεγονότα στον κόσμο και να οργανώνουν την καθημερινή τους ζωή.

Η ταξινόμηση κειμένων και η κατηγοριοποίηση ειδήσεων παίζουν σημαντικό ρόλο στην ανάλυση και οργάνωση της πληθώρας των πληροφοριών που διατίθενται. Αυτή η διαδικασία επιτρέπει στους ανθρώπους να εντοπίζουν εύκολα και να αποκτούν

πρόσβαση σε άρθρα που τους ενδιαφέρουν, βελτιώνοντας έτσι την εμπειρία τους κατά την κατανάλωση ειδήσεων.

Μέρος Β: Το προτεινόμενο πληροφοριακό σύστημα καταγραφής ειδησεογραφικών γεγονότων στα Ελληνικά

5 Ανάλυση και σχεδίαση του νέου συστήματος

5.1 Εισαγωγή

Μια βάση δεδομένων ειδήσεων προσφέρει μια οργανωμένη και προσβάσιμη αποθήκη πληροφοριών που εξυπηρετεί ένα ευρύ φάσμα σκοπών, από ιστορική ανάλυση έως αναφορές και έρευνα σε πραγματικό χρόνο. Εξουσιοδοτεί τους χρήστες με εργαλεία για την εξαγωγή πληροφοριών, την επαλήθευση πληροφοριών και τη συμβολή στην τεκμηριωμένη λήψη αποφάσεων σε διάφορους τομείς.

5.2 Συνοπτική περιγραφή του πληροφοριακού συστήματος

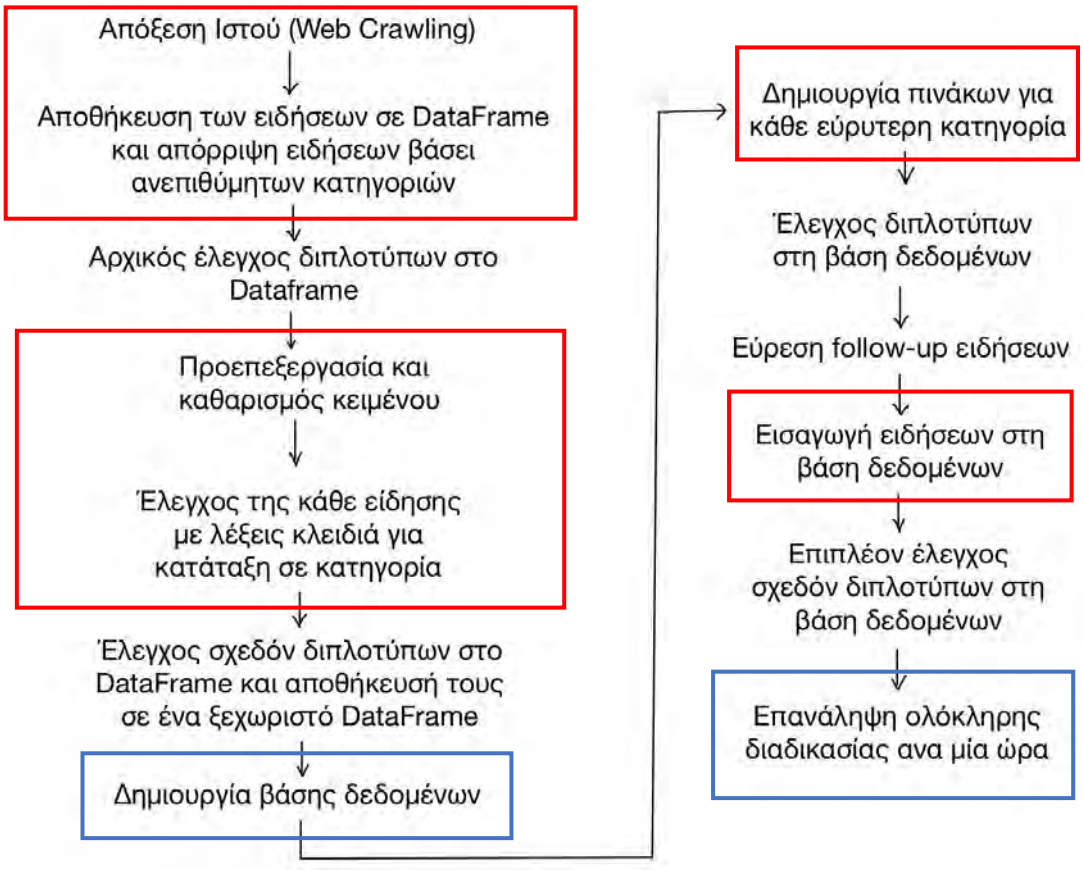
Στο δυναμικό τοπίο της ενημέρωσης και της δημοσιογραφίας, η δημιουργία μιας ισχυρής βάσης δεδομένων ειδήσεων απαιτεί μια σχολαστική και πολλαπλών βημάτων προσέγγιση. Αυτή η διαδικασία διασφαλίζει τη συλλογή, την οργάνωση και τη συνεχή βελτίωση των ειδήσεων για να παρέχει μια πολύτιμη πηγή για ερευνητές, δημοσιογράφους και αναλυτές. Η διαδικασία δημιουργίας της βάσης δεδομένων ειδησεογραφικών γεγονότων περιλαμβάνει επιγραμματικά:

1. **Απόξεση Ιστού (Web Crawling):** Η διαδικασία ξεκινά με την αυτοματοποιημένη απόξεση ιστού, όπου συλλέγονται συστηματικά άρθρα ειδήσεων από διαφορετικές ειδησεογραφικές πηγές από την ελληνική ψηφιακή σφαίρα.
2. **Αποθήκευση των ειδήσεων σε DataFrame και απόρριψη ειδήσεων βάσει ανεπιθύμητων κατηγοριών:** Οι συλλεγόμενες ειδήσεις δομούνται και αποθηκεύονται σε ένα DataFrame, επιτρέποντας την αποτελεσματική επεξεργασία και ανάλυση δεδομένων. Μια διαδικασία φιλτραρίσματος εξαλείφει τις ειδήσεις που βασίζονται σε ανεπιθύμητες κατηγορίες, διασφαλίζοντας ότι η βάση δεδομένων εστιάζει σε σχετικό και ουσιαστικό περιεχόμενο.
3. **Αρχικός έλεγχος διπλότυπων στο DataFrame**
4. **Προεπεξεργασία και καθαρισμός κειμένου**: Για να βελτιωθεί η αναγνωσιμότητα και να διευκολυνθεί η ανάλυση, το κείμενο του άρθρου

ειδήσεων υποβάλλεται σε προεπεξεργασία και καθαρισμό, βελτιστοποιώντας το για τα επόμενα στάδια.

5. **Έλεγχος της κάθε είδησης με λέξεις-κλειδιά για κατάταξη σε κατηγορία**: Χρησιμοποιώντας προηγμένες αναζητήσεις λέξεων-κλειδιών, τα άρθρα ειδήσεων κατηγοριοποιούνται σε ευρύτερες κατηγορίες ανάλογα με το θέμα που παρουσιάζουν, ενισχύοντας τις δυνατότητες πλοήγησης και έρευνας της βάσης δεδομένων.
6. **Έλεγχος σχεδόν διπλότυπων στο DataFrame και αποθήκευση τους σε ένα ξεχωριστό DataFrame**
7. **Δημιουργία βάσης δεδομένων**: Δημιουργείται η βάση δεδομένων ειδήσεων, που λειτουργεί ως δομημένο αποθετήριο για την οργάνωση των ειδησεογραφικών γεγονότων.
8. **Δημιουργία πινάκων για κάθε ευρύτερη κατηγορία**: Για να διευκολυνθεί η περιήγηση, δημιουργούνται μεμονωμένοι πίνακες στη βάση δεδομένων, ο καθένας αφιερωμένος σε μία ευρύτερη κατηγορία ειδήσεων.
9. **Έλεγχος διπλότυπων στη βάση δεδομένων**
10. **Εύρεση follow-up ειδήσεων**
11. **Εισαγωγή ειδήσεων στη βάση δεδομένων**: Οι ειδήσεις εισάγονται συστηματικά στη βάση δεδομένων.
12. **Επιπλέον έλεγχος σχεδόν διπλότυπων στη βάση δεδομένων**
13. **Επανάληψη ολόκληρης της διαδικασίας ανά μία ώρα**: Η διαδικασία δεν ολοκληρώνεται εδώ. Γίνεται ένας βρόχος που επαναλαμβάνεται κάθε ώρα. Αυτό διασφαλίζει ότι η βάση δεδομένων παραμένει ενημερωμένη, καταγράφοντας τη συνεχώς εξελισσόμενη φύση των ειδήσεων.

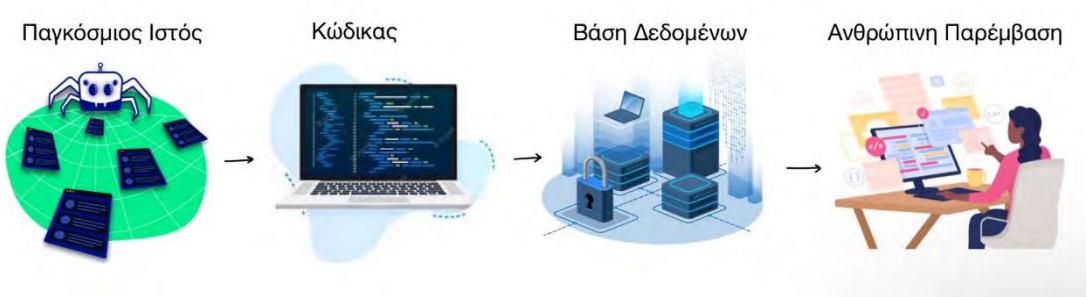
Παρακάτω εμφανίζεται ένα διάγραμμα για την απεικόνιση των βημάτων που αναφέρθηκαν. Με κόκκινο χρώμα εμφανίζονται οι διαδικασίες που υλοποιούνται στην παρούσα πτυχιακή εργασία, με μπλε χρώμα έχουν τονιστεί οι διαδικασίες που είναι κοινές και με την εργασία της Ρ. Μάντζαρη [M23] και οι υπόλοιπες διαδικασίες περιγράφονται στην εργασία [M23].



Εικόνα 14: Διάγραμμα σχεδίασης της ροής του προτεινόμενου πληροφοριακού συστήματος

5.3 Τρόπος λειτουργίας Πληροφοριακού Συστήματος

Τα Πληροφοριακά Συστήματα υποστηρίζουν τις οργανωτικές διαδικασίες και παρέχουν ένα πλαίσιο για τη συλλογή, επεξεργασία, αποθήκευση και μετατροπή των δεδομένων σε χρήσιμες πληροφορίες. Είναι η δυναμική αλληλεπίδραση της τεχνολογίας, των ανθρώπων και των διαδικασιών που συνεργάζονται για να διευκολύνουν τις αποτελεσματικές λειτουργίες, την τεκμηριωμένη λήψη αποφάσεων και τη στρατηγική ανάπτυξη.



Εικόνα 15: Τρόπος λειτουργίας Πληροφοριακού Συστήματος. Πηγές: www.istockphoto.com , www.freepik.com , https://www.scraping-bot.io/how-to-build-a-web-crawler/

Ξεκινώντας από τον Παγκόσμιο Ιστό, προχωρώντας στον κώδικα, ακολουθούμενος από τη βάση δεδομένων και, τέλος, περιλαμβάνοντας ανθρώπινη παρέμβαση για τη διασφάλιση της ακρίβειας των πληροφοριών. Αυτή η διαδικασία αντιπροσωπεύει μια ημιαυτόματη (semi-automatic) ροή εργασίας του συστήματος πληροφοριών.

Ο Παγκόσμιος Ιστός είναι το παγκόσμιο σύστημα διασυνδεδεμένων εγγράφων και πόρων που είναι προσβάσιμα μέσω του Διαδικτύου. Λειτουργεί ως μια τεράστια αποθήκη πληροφοριών που καλύπτει ένα ευρύ φάσμα θεμάτων. Οι χρήστες μπορούν να έχουν πρόσβαση σε ιστότοπους, ιστοσελίδες και περιεχόμενο πολυμέσων για τη συλλογή πληροφοριών.

Το κομμάτι του κώδικα αναφέρεται συνήθως στην παραγωγή βημάτων σε μια γλώσσα κατανοητή για τον υπολογιστή προκειμένου να δημιουργηθούν διαφόρων τύπων εφαρμογές (λογισμικού, λειτουργιών). Για την δημιουργία, λοιπόν, ενός πληροφοριακού συστήματος, οι επιστήμονες της πληροφορικής, και συγκεκριμένα οι προγραμματιστές, υλοποιούν διαδικτυακές εφαρμογές και εργαλεία για τη συλλογή, επεξεργασία και παρουσίαση πληροφοριών από τον Παγκόσμιο Ιστό στους χρήστες.

Οι βάσεις δεδομένων είναι δομημένα αποθετήρια για την αποθήκευση και τη διαχείριση δεδομένων. Διαδραματίζουν κρίσιμο ρόλο στα πληροφοριακά συστήματα παρέχοντας έναν δομημένο και οργανωμένο τρόπο αποθήκευσης και ανάκτησης δεδομένων. Όταν οι πληροφορίες συλλέγονται από τον Ιστό και υποβάλλονται σε επεξεργασία μέσω κώδικα, οδηγούνται σε βάσεις δεδομένων για αποτελεσματική αποθήκευση και ανάκτηση.

Πολλές φορές, τα δεδομένα είναι πολύπλοκα και διφορούμενα, απαιτώντας ανθρώπινη κρίση και παρέμβαση. Οι άνθρωποι έχουν βασικό ρόλο στη εξασφάλιση της ακρίβειας, της συνάφειας και του πλαισίου των δεδομένων. Αναλαμβάνουν ενέργειες, όπως η επικύρωση, ο καθαρισμός των δεδομένων και η διασφάλιση της ποιότητας, καθώς και υποκειμενικές κρίσεις για την ερμηνεία πληροφοριών που ενδέχεται να μην είναι πλήρως κατανοητές από τις αυτοματοποιημένες διαδικασίες. Η ανθρώπινη επιβεβαίωση των δεδομένων αποτελεί σημαντικό στοιχείο του ποιοτικού ελέγχου και διασφαλίζει την ακρίβεια και τη συνάφεια των πληροφοριών που παρέχονται στους χρήστες.

Συμπερασματικά, μια ημι-αυτοματοποιημένη προσέγγιση που συνδυάζει την αποτελεσματικότητα της αυτοματοποίησης με την ανθρώπινη κριτική σκέψη και κρίση μπορεί να δημιουργήσει μια ισχυρή και αξιόπιστη βάση δεδομένων ειδήσεων. Μπορεί να αξιοποιήσει την τεχνολογία για την αντιμετώπιση μεγάλων εργασιών, παρέχοντας ταυτόχρονα την ακρίβεια, την ποιότητα και την κατανόηση του πλαισίου που μόνο η ανθρώπινη παρέμβαση μπορεί να προσφέρει.

5.4 Ανάλυση modules, πακέτων, βιβλιοθηκών και drivers του κώδικα

Προτού ξεκινήσει η ανάλυση της συγκεκριμένης ενότητας, απαραίτητη είναι η ανάλυση κάποιων εννοιών. Συγκεκριμένα, θα πρέπει να ξεκαθαριστούν οι όροι scripts, modules, πακέτα (packages), βιβλιοθήκες (libraries) και drivers.

- Ένα script είναι ένα αρχείο Python που προορίζεται για άμεση εκτέλεση. Όταν αυτό εκτελείται, πρέπει να πραγματοποιεί μια λειτουργία. Αυτό σημαίνει, ότι τα scripts θα περιέχουν συχνά κώδικα γραμμένο εκτός του πεδίου εφαρμογής οποιονδήποτε κλάσεων ή συναρτήσεων.
- Ένα module είναι ένα αρχείο Python που προορίζεται να εισαχθεί σε scripts ή άλλα modules. Συχνά ορίζει μέλη, όπως κλάσεις, συναρτήσεις και μεταβλητές που προορίζονται να χρησιμοποιηθούν σε άλλα αρχεία που το εισάγουν, χρησιμοποιώντας την φράση import.
- Ένα πακέτο είναι μια συλλογή σχετικών modules που συνεργάζονται για να παρέχουν ορισμένες λειτουργίες. Αυτά τα modules περιέχονται σε ένα φάκελο και μπορούν να εισαχθούν όπως όλα τα άλλα modules. Αυτός ο φάκελος, θα περιέχει συχνά ένα ειδικό αρχείο __init__ που λέει στην Python ότι είναι ένα πακέτο, που ενδεχομένως περιέχει περισσότερες ενότητες που είναι ένθετες σε υποφακέλους.
- Μια βιβλιοθήκη είναι ένας όρος ομπρέλα που σημαίνει "μια δέσμη κώδικα". Μπορεί να έχει δεκάδες ή και εκατοντάδες μεμονωμένα modules που μπορούν να παρέχουν ένα ευρύ φάσμα λειτουργιών [C20].
- Ένας driver είναι ένα σύνολο αρχείων που λέει σε ένα κομμάτι του υλικού πώς να λειτουργεί επικοινωνώντας με το λειτουργικό σύστημα ενός υπολογιστή [K19].

Για την πραγματοποίηση της υλοποίησης του κώδικα ήταν απαραίτητη πληθώρα βιβλιοθηκών, πακέτων, modules και drivers ανάλογα με τις επιθυμητές

λειτουργίες υλοποίησης. Βέβαια, θα πρέπει να γίνει η εγκατάστασή τους σε συγκεκριμένο φάκελο προκειμένου να εκτελεστούν σωστά. Αν στο τρέχοντα φάκελο (.) βρίσκεται το αρχείο με τον κώδικα, τότε οι βιβλιοθήκες, τα πακέτα, τα modules και οι drivers θα πρέπει να εγκαθίστανται στο μονοπάτι *./venv/Scripts*.

5.4.1 Modules

Feed parser: είναι ένα module στην Python για λήψη και ανάλυση κοινοπρακτικών ροών. Μπορεί να χειριστεί ροές RSS, Atom και CDF.

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install feedparser* [PM15].

Time: παρέχει διάφορες λειτουργίες που σχετίζονται με το χρόνο.

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install python-time* [PSF23a].

Unicode: παρέχει πρόσβαση στη βάση δεδομένων χαρακτήρων Unicode (UCD) που ορίζει ιδιότητες χαρακτήρων για όλους τους χαρακτήρες Unicode

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install Unidecode* [PSF23b].

Openpyxl: είναι μια βιβλιοθήκη της Python για την ανάγνωση/εγγραφή Excel xlsx/xlsm/xltx/xltm αρχείων.

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install openpyxl* [XG23].

5.4.2 Πακέτα

Pandas: είναι ένα πακέτο Python που παρέχει γρήγορες, ευέλικτες και εκφραστικές δομές δεδομένων που έχουν σχεδιαστεί για να κάνουν την εργασία με "σχεσιακά" ή "επισημασμένα (labeled)" δεδομένα τόσο εύκολη όσο και διαισθητική. Στόχος του, το να είναι το θεμελιώδες δομικό στοιχείο υψηλού επιπέδου για την πραγματοποίηση πρακτικής ανάλυσης δεδομένων πραγματικού κόσμου στην Python.

Για την φόρτωση του συγκεκριμένου πακέτου απαραίτητη είναι η χρήση της εντολής στο command line *pip install pandas* [TPDT23].

NLTK: To Natural Language Toolkit (NLTK) είναι ένα πακέτο Python για επεξεργασία φυσικής γλώσσας. Περιλαμβάνει την ανάλυση, την ποσοτικοποίηση, την κατανόηση και την εξαγωγή νοήματος από φυσικές γλώσσες.

Για την φόρτωση του συγκεκριμένου πακέτου απαραίτητη είναι η χρήση της εντολής στο command line *pip install nltk* [NltkT23].

Wheel: είναι μια ενσωματωμένη μορφή πακέτου για Python. Το wheel είναι ένα αρχείο μορφής ZIP με ένα ειδικά διαμορφωμένο όνομα αρχείου και την επέκταση *.whl*. Έχει σχεδιαστεί για να περιέχει όλα τα αρχεία για μια εγκατάσταση συμβατή με PEP 376 με τρόπο που να είναι πολύ κοντά στη μορφή του δίσκου. Πολλά πακέτα θα εγκατασταθούν σωστά μόνο με το βήμα "Unpack" (απλώς εξαγωγή του αρχείου στο sys.path) και το unpack αρχείο διατηρεί αρκετές πληροφορίες για "Spread" (αντιγραφή δεδομένων και σεναρίων στις τελικές τοποθεσίες τους) οποιαδήποτε μεταγενέστερη στιγμή.

Για την φόρτωση του συγκεκριμένου πακέτου απαραίτητη είναι η χρήση της εντολής στο command line *pip install -U pip setuptools wheel* [H14].

5.4.3 Βιβλιοθήκες

Requests: είναι μια απλή βιβλιοθήκη HTTP, συγκεκριμένα επιτρέπει την αποστολή αιτημάτων HTTP χρησιμοποιώντας Python.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install requests* [RD23].

Beautiful Soup: είναι μια βιβλιοθήκη που διευκολύνει την απόξεση πληροφοριών από ιστοσελίδες, που συνήθως είναι από αρχεία HTML και XML.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install beautifulsoup4* [R23].

Pytz: είναι μια βιβλιοθήκη που φέρνει τη βάση δεδομένων Olson tz στην Python και έτσι υποστηρίζει σχεδόν όλες τις ζώνες ώρας. Εξυπηρετεί τις λειτουργίες μετατροπής ημερομηνίας-ώρας. Επιτρέπει τους υπολογισμούς ζώνης ώρας στις εφαρμογές Python και επιτρέπει επίσης την δημιουργία στιγμιοτύπων ημερομηνίας με επίγνωση της ζώνης ώρας.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install pytz* [D22].

Greek-Stemmer: Η βιβλιοθήκη stemming της Python αφαιρεί το επίθημα μιας λέξης σύμφωνα με ένα σύνολο αλγορίθμων που βασίζονται σε κανόνες. Ο αλγόριθμος αναπτύσσεται σύμφωνα με τους γραμματικούς κανόνες της νεοελληνικής γλώσσας. Μια εκτεταμένη τεκμηρίωση της διαδικασίας αφαίρεσης, καθώς και μια σύντομη αξιολόγηση του συστήματος δείχνει την ακρίβεια του αλγορίθμου που λειτουργεί με καλύτερη απόδοση από άλλους παλιούς αλγόριθμους βασικής αρχής για την ελληνική γλώσσα δίνοντας 99,4 τοις εκατό σωστά αποτελέσματα σε ένα σύνολο δεδομένων 5000 λέξεων.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install greek-stemmer* [P21],[L17].

Itertools: Η βιβλιοθήκη itertools της Python μπορεί να συνθέσει κομψές λύσεις για μια ποικιλία προβλημάτων με τις λειτουργίες που παρέχει. Με την more-itertools γίνεται συλλογή επιπρόσθετων δομικών στοιχείων και ρουτίνες για την εργασία με Python iterables.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install more-itertools* [Ro23].

SQLAlchemy: είναι το kit εργαλείων Python SQL και το Object Relational Mapper που δίνει στους προγραμματιστές εφαρμογών την πλήρη ισχύ και την ευελιξία της SQL. Με άλλα λόγια είναι μια αφηρημένη βιβλιοθήκη βάσεων δεδομένων.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install SQLAlchemy* [B23b],[Alc23].

Os-Sys: είναι μια μεγάλη βιβλιοθήκη με πολλά χρήσιμα εργαλεία, όπως την διόρθωση σφαλμάτων, την δημιουργία γρηγορότερων συναρτήσεων και κλάσεων κτλ.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install os_sys* [L19].

SpaCy: είναι μια βιβλιοθήκη ανοιχτού κώδικα με πολλές ενσωματωμένες δυνατότητες. Γίνεται όλο και πιο δημοφιλές για την επεξεργασία και την ανάλυση δεδομένων στον τομέα του NLP.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install -U spacy* [S23].

Επιπρόσθετα, γίνεται η εγκατάσταση ενός pipeline προκειμένου να είναι εφικτή η επεξεργασία φυσικής γλώσσας στα ελληνικά. Συγκεκριμένα αυτό επιτυγχάνεται με την εντολή στο command line *python -m spacy download el_core_news_sm*. [Spa23]. Βέβαια, θα πρέπει να εκτελεστούν με σειρά οι εντολές *pip install -U pip setuptools wheel / pip install -U spacy / python -m spacy download el_core_news_sm.*

5.4.4 Drivers

MySQL Connector: επιτρέπει στα προγράμματα της python να έχουν πρόσβαση σε βάσεις δεδομένων MySQL, χρησιμοποιώντας ένα API που είναι συμβατό με τη Python Database.

Για την φόρτωση του συγκεκριμένου driver απαραίτητη είναι η χρήση της εντολής στο command line *pip install mysql-connector-python* [Ora23a].

5.5 Επίλογος

Τέλος, σχεδιάστηκε και αναλύθηκε το σύστημα πληροφοριών του έργου. Με άλλα λόγια, περιγράφηκε εν συντομία ο τρόπος λειτουργίας του συστήματος αυτού από το διαδίκτυο μέχρι το τελικό προϊόν και τα βήματα που ακολουθήθηκαν από τον κώδικα. Επιπλέον, παρουσιάστηκαν οι βιβλιοθήκες, τα πακέτα, τα modules και οι drivers που είναι απαραίτητα για την λειτουργικότητα του κώδικα.

6 Συλλογή, επεξεργασία και αποθήκευση περιεχομένου ειδησεογραφικών άρθρων

6.1 Συλλογή ειδησεογραφικών άρθρων και αποθήκευσή τους σε dataframe

6.1.1 Εισαγωγή

Η συλλογή των ειδησεογραφικών άρθρων πραγματοποιείται μέσω διαδικτυακών ειδησεογραφικών μέσων ενημέρωσης. Τα διαδικτυακά ειδησεογραφικά μέσα, γνωστά και ως διαδικτυακοί ιστότοποι ειδήσεων ή ψηφιακές πλατφόρμες ειδήσεων, είναι οργανισμοί μέσων ενημέρωσης που βασίζονται στο Διαδίκτυο και παρέχουν ειδήσεις και πληροφορίες στο κοινό τους μέσω ψηφιακών καναλιών. Για την διαχείριση των πληροφοριών από τα ειδησεογραφικά μέσα ενημέρωσης είναι χρήσιμη η αποθήκευσή τους σε μία δομή δεδομένων που προσφέρει η Python, η οποία ονομάζεται DataFrame.

6.1.2 Συλλογή ειδησεογραφικών άρθρων

Τα διαδικτυακά ειδησεογραφικά μέσα προσφέρουν διάφορα πλεονεκτήματα σε σχέση με τα παραδοσιακά έντυπα μέσα ενημέρωσης και την τηλεοπτική μετάδοση. Οποιοσδήποτε με πρόσβαση στο Διαδίκτυο μπορεί να επισκεφτεί ειδησεογραφικούς ιστότοπους από οπουδήποτε και ανά πάσα στιγμή, καθιστώντας τις ειδήσεις άμεσα διαθέσιμες σε ένα παγκόσμιο κοινό. Επιπλέον, τα διαδικτυακά ειδησεογραφικά πρακτορεία μπορούν να παρέχουν ενημερώσεις σχετικά με τις έκτακτες ειδήσεις και την εξέλιξή τους, κρατώντας το κοινό ενημερωμένο σε πραγματικό χρόνο.

Είναι αναγκαία η επιλογή αξιόπιστων διαδικτυακών μέσων ενημέρωσης και ιστότοπους ειδήσεων από τους οποίους θα πραγματοποιηθεί η συλλογή των δεδομένων ειδήσεων για τη βάση δεδομένων. Αυτό είναι ένα κρίσιμο βήμα, καθώς η αξιοπιστία των πηγών ειδήσεων επηρεάζει άμεσα την ποιότητα και την αξιοπιστία των πληροφοριών.

Ορισμένα βασικά χαρακτηριστικά που πρέπει να κατέχει μια πηγή ειδήσεων είναι τα εξής:

- Καθιερωμένα και αξιόπιστα ειδησεογραφικά πρακτορεία που έχουν ιστορικό παροχής ακριβούς και αμερόληπτης κάλυψης ειδήσεων. Πηγές που τηρούν υψηλά δημοσιογραφικά πρότυπα και έχουν ισχυρή φήμη για πραγματικές αναφορές.

- Σαφείς συντακτικές οδηγίες και πολιτικές που δίνουν προτεραιότητα στην ακρίβεια και την αντικειμενικότητα στις αναφορές. Αποφυγή πηγών που είναι γνωστές για εντυπωσιασμούς ή μεροληπτικές αναφορές.
- Επαλήθευση πληροφοριών με πολλές πηγές και παροχή αποδεικτικών στοιχείων ή συνδέσμους προς τις πηγές τους όποτε είναι δυνατόν.
- Κάλυψη διαφορετικών θεμάτων και περιοχών ώστε να δημιουργηθεί μια ολοκληρωμένη βάση δεδομένων που αντιπροσωπεύει διάφορες πτυχές ενημέρωσης.

Επιλέγοντας αξιόπιστες πηγές ειδήσεων, μπορεί να δημιουργηθεί μια αξιόπιστη βάση δεδομένων με άρθρα ειδήσεων που μπορεί να χρησιμοποιηθεί για ανάλυση και έρευνα.

Αν και η ίδια η απόξεση δεδομένων δεν είναι εγγενώς κακόβουλη, ορισμένες πρακτικές μπορεί να οδηγήσουν σε αρνητικές συνέπειες. Πρώτον, η απόξεση ειδησεογραφικών άρθρων χωρίς άδεια μπορεί να παραβιάζει τους νόμους περί πνευματικών δικαιωμάτων και τα δικαιώματα πνευματικής ιδιοκτησίας, οδηγώντας σε νομικές επιπτώσεις. Επιπλέον, ορισμένα μέσα ενημέρωσης ενδέχεται να εφαρμόσουν μέτρα ασφαλείας για να αποτρέψουν την απόξεση και οι προσπάθειες παράκαμψης αυτών των μέτρων θα μπορούσαν να θεωρηθούν ανήθικες και ενδέχεται να οδηγήσουν σε νομικές ενέργειες. Για να αντιμετωπιστούν αυτές οι ανησυχίες, είναι σημαντική η συμμόρφωση με τις δεοντολογικές οδηγίες και η αναζήτηση της κατάλληλης εξουσιοδότησης προτού αφαιρεθεί περιεχόμενο ειδήσεων από διαδικτυακά μέσα ενημέρωσης.

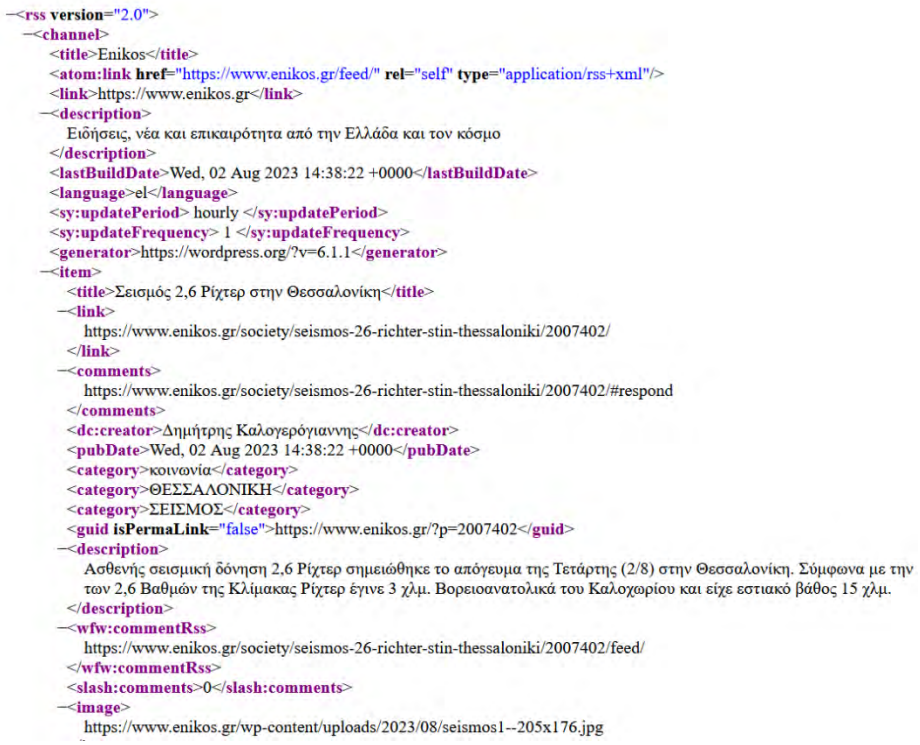
Σύμφωνα με τον ελληνικό κατάλογο ιστοσελίδων όσο αφορά τις ειδήσεις και τα Μέσα Μαζικής Ενημέρωσης για το 2023, η κατάταξη των διαδικτυακών πηγών ενημέρωσης είναι οι εξής:

1. Protothema.gr
2. Newsit.gr
3. In.gr
4. Iefimerida.gr
5. Pronews.gr
6. News247.gr

7. Newsbomb.gr
8. Zougla.gr
9. Makeleio.gr
10. Kathimerini.gr
11. Enikos.gr
12. Dimokratia.gr
13. Tovima.gr
14. Pagenews.gr
15. Topontiki.gr

Να σημειωθεί ότι οι παραπάνω διαδικτυακές πηγές ενημέρωσης είναι δίχως εκείνων που αφορούν μόνο αθλητικά, τηλεοπτικά κανάλια και γενικά πηγές που αφορούν μόνο ένα θέμα ενημέρωσης όπως lifestyle. Ώστε η βάση δεδομένων να είναι αμερόληπτη και αντιπροσωπευτική, αποφασίστηκε να απορριφθούν ειδησεογραφικές πηγές που ανήκουν σε ίδιους εκδοτικούς οίκους για να υπάρχει η μεγαλύτερη δυνατή κάλυψη διαφορετικών απόψεων. Η επιλογή του ποια πηγή θα μείνει ανάμεσα σε αυτές του ίδιου οίκου ήταν τυχαία.

Επιπλέον, η τελική επιλογή των ειδησεογραφικών πηγών έγινε με βάση τα δομικά στοιχεία του κάθε ιστοτόπου, ώστε να είναι συμβατά με τη διαδικασία απόξεσης των ειδήσεων. Ελέγχθηκε η ύπαρξη RSS feed της κάθε διαδικτυακής πηγής όπως και η δομή του κάθε feed, για να γίνει κατάλληλη επιλογή εντολών στον κώδικα των συναρτήσεων, ανάλογα με την ετικέτα HTML που χρησιμοποιείται σε κάθε πεδίο.



Εικόνα 16: Παράδειγμα RSS feed του enikos.gr

Με την εκτέλεση του κώδικα, αντιμετωπίστηκαν ορισμένα προβλήματα που περιόρισαν ακόμη περισσότερο την τελική λίστα των ειδησεογραφικών πηγών. Ορισμένες ιστοσελίδες ακολουθούν μέτρα ασφαλείας ώστε να αποτρέψουν την απόξεση τους, υπό την υποψία κακόβουλων προθέσεων, με αποτέλεσμα τον τερματισμό του κώδικα έπειτα από ορισμένες επαναλήψεις. Εφόσον οι προσπάθειες παράκαμψης αυτών των μέτρων θα μπορούσαν να θεωρηθούν ανήθικες και ενδέχεται να οδηγήσουν σε νομικές ενέργειες, αποφασίστηκε οι συγκεκριμένες πηγές να απορριφθούν.

Με βάση τα παραπάνω, η τελική λίστα των ειδησεογραφικών πηγών τροποποιήθηκε ως εξής:

1. Enikos.gr
2. Dimokratia.gr
3. Tovima.gr
4. Pagenews.gr
5. Naftemporiki.gr

Ακολουθεί η συνάρτηση που χρησιμοποιείται για την απόξεση της ιστοσελίδας enikos.gr, ως παράδειγμα περιγραφής της διαδικασίας που ακολουθήθηκε. Παρόμοια

διαδικασία ακολουθούν οι συναρτήσεις των προαναφερθέντων ειδησεογραφικών πηγών με τις κατάλληλες αλλαγές στον κώδικα.

```
# Εξαγωγή ειδήσεων από το enikos.gr
def read_enikos_rss():
    # URL προς απόξεση
    url_to_scrape = "https://www.enikos.gr/feed/"
    NewsFeed = feedparser.parse(url_to_scrape)

    news_items = []
    for entry in NewsFeed.entries:
        news_item = {}
        news_item['title'] = entry.title
        news_item['category'] = entry.category
        news_item['description'] = entry.description
        news_item['link'] = entry.link
        datetime_rss = parser.parse(entry.published)
        news_item['pubDate'] = datetime_rss.astimezone(timeZ_At).strftime("%d-%m-%Y %H:%M:%S")

        #  Ανάγνωση του περιεχομένου των ειδήσεων
        news = requests.get(entry.link)
        news_content = news.content
        soup_news = BeautifulSoup(news_content, 'html.parser')
        for script in soup_news(['script']):
            script.decompose()  # αφαίρεση ετικετών script και του περιεχομένου τους

        b = soup_news.find_all('div', class_='articletext')
        try:
            x = b[0].get_text()
        except IndexError:
            x = ""

        # Αφαίρεση κενών γραμμών από το περιεχόμενο
        lines = x.split("\n")
        non_empty_lines = [line for line in lines if line.strip() != ""]
        text_without_empty_lines = ""
        for line in non_empty_lines:
            text_without_empty_lines += line + " "
        news_item['content'] = re.sub(r"http\S+", "", text_without_empty_lines)

        news_items.append(news_item)
    return news_items
```

Εικόνα 17: Κώδικας απόξεσης ειδήσεων από τον ειδησεογραφικό ιστότοπο enikos.gr

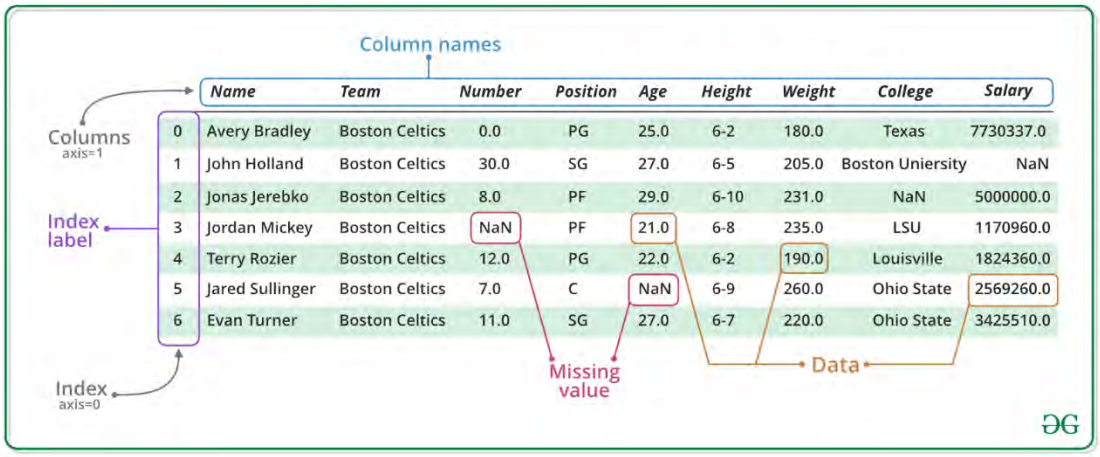
Για κάθε άρθρο ειδήσεων, δηλαδή για κάθε εγγραφή, που βρίσκεται στο feed του κάθε ιστοτόπου, γίνεται αποθήκευση του τίτλου, της κατηγορίας που έχει ανατεθεί από την εκάστοτε ιστοσελίδα, την περιγραφή του άρθρου, την ηλεκτρονική διεύθυνση (link), την ημερομηνία και ώρα δημοσίευσης και τέλος, το κύριο σώμα – περιεχόμενο της είδησης. Επιπλέον, πραγματοποιείται αφαίρεση κενών γραμμών από το περιεχόμενο, αν υπάρχουν, για εξοικονόμηση χώρου αποθήκευσης. Κάθε είδηση αποθηκεύεται στη λίστα news_items.

Το σημείο του κώδικα της συνάρτησης που αλλάζει ανά ειδησεογραφικό κανάλι, είναι κυρίως το πεδίο που περιέχει το περιεχόμενο (content) της κάθε είδησης.

Αυτό συμβαίνει, διότι, είτε το κάθε ειδησεογραφικό έχει άλλη ετικέτα HTML για να ορίσει το περιεχόμενο.

6.1.3 Αποθήκευση ειδησεογραφικών άρθρων σε DataFrame

Στην Python, ένα DataFrame αναφέρεται σε μια δισδιάστατη δομή δεδομένων που παρέχεται από τη δημοφιλή βιβλιοθήκη που ονομάζεται Pandas. Είναι μια από τις πιο ευρέως χρησιμοποιούμενες δομές δεδομένων για χειρισμό και ανάλυση δεδομένων. Ένα DataFrame μπορεί να θεωρηθεί ως μια δομή δεδομένων σαν πίνακα, παρόμοια με ένα υπολογιστικό φύλλο ή έναν πίνακα SQL. Αποτελείται από σειρές και στήλες, όπου κάθε στήλη μπορεί να έχει έναν συγκεκριμένο τύπο δεδομένων, όπως αριθμητικό, συμβολοσειρά. Τα DataFrames επιτρέπουν την αποθήκευση και τον χειρισμό δομημένων δεδομένων αποτελεσματικά. Ένα Pandas DataFrame μπορεί να δημιουργηθεί από λίστες, λεξικά και από λίστα λεξικών.



Εικόνα 18: Παράδειγμα μορφής DataFrame που περιέχει παίκτες ποδοσφαίρου και τα χαρακτηριστικά τους. Πήγη: geeksforgeeks.org

Τα DataFrames προσφέρουν πολυάριθμες λειτουργίες, όπως φιλτράρισμα, συγχώνευση, ομαδοποίηση και πολλά άλλα, καθιστώντας εύκολη την αποτελεσματική εκτέλεση διαφόρων λειτουργιών ανάλυσης δεδομένων. Επιπλέον, μπορούν εύκολα να χειριστούν δεδομένα που λείπουν και να χειριστούν διαφορετικούς τύπους δεδομένων εντός της ίδιας δομής, γεγονός που τα καθιστά εξαιρετικά ευέλικτα για εργασίες χειρισμού δεδομένων.

Ακολουθεί το απόσπασμα κώδικα με τη δημιουργία του DataFrame με τα δεδομένα που συλλέχθηκαν από την ιστοσελίδα enikos.gr. Κάθε χαρακτηριστικό της είδησης που αποθηκεύτηκε προηγουμένως στη λίστα news_items αποθηκεύεται σε

διαφορετική στήλη του DataFrame. Παρόμοια διαδικασία ακολουθούν και οι υπόλοιπες ειδησεογραφικές πηγές, με τις κατάλληλες αλλαγές στον κώδικα.

```
# Δημιουργία dataframe df5 ειδήσεων του enikos.gr
df = pd.DataFrame(read_enikos_rss(), columns=['title', 'category', 'description', 'link', 'content', 'pubDate'])
```

Εικόνα 19: Δημιουργία DataFrame για το ειδησεογραφικό enikos.gr

Αφότου πραγματοποιηθεί η αποθήκευση όλων των δεδομένων της κάθε ιστοσελίδας σε διαφορετικό DataFrame, γίνεται η μεταξύ τους συνένωση σε ένα συνολικό DataFrame, το οποίο έπειτα ταξινομείται με βάση την ημερομηνία και ώρα δημοσίευσης της κάθε εγγραφής.

```
# συνένωση των dataframes df1, df2, df3, df4, df5, df6
frames = [df1, df2, df3, df4, df5, df6]
df = pd.concat(frames)
df.reset_index(drop=True, inplace=True)
```

Εικόνα 20: Συνένωση όλων των DataFrame από τις ειδησεογραφικές πηγές.

Είναι απαραίτητο να σημειωθεί ότι το συνολικό DataFrame ελέγχεται με βάση τη κατηγορία της εκάστοτε είδησης και απορρίπτονται ειδήσεις που αφορούν:

- Lifestyle
- Μαγειρικό περιεχόμενο
- Τηλεοπτικά προγράμματα
- Ειδήσεις με άγνωστη κατηγορία
- Άλλες ειδήσεις με ανούσιο περιεχόμενο

6.1.4 Επίλογος

Συμπερασματικά, η επιλογή των ειδησεογραφικών μέσων ενημέρωσης είναι κρίσιμης σημασίας προκειμένου να συλλεχθούν έγκυρες ειδησεογραφικές πληροφορίες γεγονότων. Βέβαια, είναι απαραίτητη μια προεπεξεργασία των άρθρων ώστε να γίνει η απόξεση συγκεκριμένων τμημάτων της είδησης από τον ιστό και η αποθήκευση τους σε DataFrame. Με αυτόν τον τρόπο, γίνεται ευκολότερη η μετέπειτα επεξεργασία των δεδομένων, για τη δημιουργία μίας αντικειμενικής, αξιόπιστης και ολοκληρωμένης βάσης δεδομένων.

6.2 Προεπεξεργασία και καθαρισμός δεδομένων ειδησεογραφικών άρθρων

6.2.1 Εισαγωγή

Η προεπεξεργασία κειμένου είναι ένα κρίσιμο βήμα στις εργασίες επεξεργασίας φυσικής γλώσσας και ανάλυσης κειμένου. Περιλαμβάνει τον καθαρισμό και την προετοιμασία δεδομένων ακατέργαστου κειμένου προτού υποβληθούν στην επιθυμητή ανάλυση. Τα δεδομένα κειμένου, ειδικά όταν προέρχονται από διάφορες πηγές, όπως

ιστότοπους, μέσα κοινωνικής δικτύωσης ή έγγραφα, μπορεί να είναι αδόμητα και να περιέχουν θόρυβο. Η προεπεξεργασία κειμένου συμβάλλει στη βελτίωση της ποιότητας των δεδομένων και στη βελτίωση της ακρίβειας και της αποτελεσματικότητας της επακόλουθης ανάλυσης.

Όπως προαναφέρθηκε στο θεωρητικό μέρος, η προεπεξεργασία των δεδομένων αποτελείται συνήθως από τα παρακάτω βήματα: τμηματοποίηση, αφαίρεση λέξεων τερματισμού, αφαίρεση θορύβου και stemming ή λημματοποίηση, με αυτή τη σειρά.

6.2.2 Προεπεξεργασία και καθαρισμός κειμένου

Τα άρθρα ειδήσεων τα οποία συναντώνται στις διαδικτυακές πηγές από τις οποίες γίνεται η συλλογή τους, περιέχουν πολλές φορές περιεχόμενο γραμμένο στην αγγλική γλώσσα, πέρα από την ελληνική. Όταν ασχολούμαστε με πολύγλωσσα δεδομένα, όπως τα αγγλικά και τα ελληνικά, η σημασία της αφαίρεσης λέξεων τερματισμού γίνεται ακόμη πιο έντονη. Καταργώντας τις άσχετες και συχνά εμφανιζόμενες λέξεις, δημιουργούμε ένα πιο καθαρό και πιο ουσιαστικό σύνολο δεδομένων, θέτοντας το υπόβαθρο για πιο ακριβή, διορατική και αποτελεσματική ανάλυση κειμένου.

Η προσαρμογή του συνόλου των λέξεων τερματισμού ώστε να περιλαμβάνει και τις δύο γλώσσες διασφαλίζει ότι η ανάλυσή μας είναι περιεκτική και προσαρμοσμένη στις συγκεκριμένες γλωσσικές αποχρώσεις του κειμένου. Επομένως, προτού πραγματοποιηθεί οποιαδήποτε ενέργεια για την προεπεξεργασία του κειμένου, είναι αναγκαία η λήψη αγγλικών και ελληνικών λέξεων τερματισμού και προσθήκης προσαρμοσμένων ελληνικών λέξεων τερματισμού που επιλέχθηκαν στα πλαίσια της πτυχιακής εργασίας. Βέβαια, δεν γίνεται κατηγοριοποίηση των αγγλικών γεγονότων, στα πλαίσια της πτυχιακής εργασίας, παρά μόνο των ελληνικών γεγονότων, απλά επειδή γίνεται συλλογή αγγλικών και ελληνικών, καλούμαστε να τις επεξεργαστούμε όλες. Επιπλέον, πραγματοποιείται ο ορισμός του εργαλείου υπεύθυνο για τη τμηματοποίηση (stemmer) συγκεκριμένα για την ελληνική γλώσσα [AZHL22].

Ακολουθεί το τμήμα κώδικα που περιγράφει τη διαδικασία ορισμού του ελληνικού stemmer **GreekStemmer()** και τη λήψη των λιστών που περιέχουν τις αγγλικές και ελληνικές λέξεις τερματισμού. Το αρχείο **extendstopwords.txt** (υπάρχει στο αρχείο Παραδοτέο-1_extendstopwords) που αναγράφεται στον κώδικα, περιέχει επιπλέον ελληνικές λέξεις τερματισμού που επιλέχθηκαν με βάση τις ανάγκες της εργασίας.

Συγκεκριμένα, πραγματοποιήθηκε επιλογή λέξεων ύστερα από ανάλυση πολυάριθμων άρθρων ειδήσεων που συλλέχθηκαν και εκτύπωση των λέξεων που τις αποτελούν, χρησιμοποιώντας τη διαδικασία της τμηματοποίησης που θα αναλυθεί παρακάτω. Το αρχείο αυτό, μετατρέπεται σε λίστα και έπειτα, συνενώνεται με τις δυο προυπάρχουσες, δημιουργώντας μια ολοκληρωμένη, τελική λίστα λέξεων τερματισμού που θα χρησιμοποιηθεί στην διαδικασία καθαρισμού του κειμένου.

```
# Χωρισμός των ειδήσεων σε λέξεις
stemmer = GreekStemmer()  # stemmer για την ελληνική γλώσσα

stopWordsGreek = stopwords.words('greek')  # Λήψη της λίστας με τις ελληνικές stop words
stopWordsEnglish = stopwords.words('english')  # Λήψη της λίστας με τις αγγλικές stop words
stopWords = stopWordsGreek + stopWordsEnglish

list_stopWords = []
with open("extendstopwords.txt", "r", encoding="utf8") as myfile:
    data = myfile.read()
    list_stopWords = data.splitlines()

stopWords.extend(list_stopWords)
```

Εικόνα 21: Λίστες λέξεων τερματισμού και ορισμός ελληνικού stemmer

Ακολουθεί η συνάρτηση **tokenize_and_stem** που χρησιμοποιείται για τη τμηματοποίηση (δημιουργία tokens), τον καθαρισμό του κειμένου και τη μετατροπή στη βασική μορφή των λέξεων.

```
# συνάρτηση για δημιουργία tokens, καθαρισμός του κειμένου και stemming
def tokenize_and_stem(str_input):
    bagOfWords = nltk.word_tokenize(str_input)  # δημιουργία tokens

    stem_sentence = []
    for token in bagOfWords:
        token = token.lower()
        # αφαίρεση κάθε τιμής που δεν ανήκει στο ελληνικό ή αγγλικό αλφάβητο
        new_token = re.sub(r'[^α-ωΑ-Ωά-ώΆ-ΏΑ-Ωa-zA-Z]+', '', token)
        # αφαίρεση των tokens μεγέθους ενός χαρακτήρα
        if len(new_token.strip()) >= 2:
            vowels = len([v for v in new_token if v in "αάεέιίηήοόυύωώaeiouy"])
            # αφαίρεση των λέξεων που περιέχουν μόνο σύμφωνα και δεν ανήκουν στα stop words
            if vowels != 0 and new_token not in stopWords:
                # μετατροπή σε κεφαλαία, αφαίρεση τόνων και stemming των λέξεων
                stem_sentence.append(stemmer.stem(ud.normalize('NFD', new_token.upper()).translate(d)))

    return stem_sentence
```

Εικόνα 22: Συνάρτηση tokenize_and_stem.

Ακολουθεί μια σύντομη εξήγηση για κάθε βήμα της συνάρτησης:

- Τμηματοποίηση: Το κείμενο εισαγωγής χωρίζεται σε μεμονωμένες λέξεις ή tokens, δημιουργώντας ένα «bag of words». Το "bag of words" (BoW) είναι μια θεμελιώδης έννοια στην επεξεργασία φυσικής γλώσσας και στην ανάλυση

κειμένου. Είναι μια αναπαράσταση δεδομένων κειμένου όπου το περιεχόμενο και η σειρά των λέξεων αγνοούνται και λαμβάνεται υπόψη μόνο η συχνότητα των λέξεων σε ένα έγγραφο ή ένα κομμάτι κειμένου. Αυτό το αρχικό βήμα αναλύει το κείμενο στα συστατικά μέρη του.

- Μετατροπή σε πεζά γράμματα: Κάθε token μετατρέπεται σε πεζά. Αυτό διασφαλίζει ότι οι λέξεις σε διαφορετικές περιπτώσεις αντιμετωπίζονται ως ίδιες, μειώνοντας τις παραλλαγές στο σύνολο δεδομένων.
- Αφαίρεση θορύβου: Οι χαρακτήρες που δεν ανήκουν στο ελληνικό ή αγγλικό αλφάβητο αφαιρούνται από κάθε token. Αυτό το βήμα καταργεί τους ειδικούς χαρακτήρες, τα σύμβολα και τους αριθμούς που ενδέχεται να μην συμβάλλουν στο γλωσσικό περιεχόμενο [BSH21].
- Ελάχιστο μήκος token: Τα διακριτικά με μήκος ενός χαρακτήρα απορρίπτονται. Αυτό αφαιρεί πολύ σύντομα και δυνητικά μη ενημερωτικά token.
- Καταμέτρηση φωνηέντων: Μετράται ο αριθμός των φωνηέντων σε κάθε token. Τα token χωρίς φωνήεντα αποτελούν λέξεις που περιέχουν μόνο σύμφωνα και αφαιρούνται, εφόσον δεν αποτελούν μέρος των προκαθορισμένων λέξεων τερματισμού.
- Αφαίρεση λέξεων τερματισμού: Η συνάρτηση ελέγχει εάν το token δεν βρίσκεται στη λίστα των λέξεων τερματισμού. Εάν δεν είναι λέξη τερματισμού, προχωρά σε περαιτέρω επεξεργασία, αντιθέτως την αγνοεί [DZ11].
- Κανονικοποίηση κειμένου: Το κάθε token μετατρέπεται σε κεφαλαία, τα διακριτικά (τονισμοί και τόνοι) αφαιρούνται χρησιμοποιώντας την κανονικοποίηση Unicode και η μετατροπή στη βασική μορφή των λέξεων εφαρμόζεται χρησιμοποιώντας ένα stemmer. Αυτό το βήμα μειώνει τις λέξεις στη ρίζα τους, κάτι που βοηθά στην ομαδοποίηση παρόμοιων λέξεων.

Τα προεπεξεργασμένα διακριτικά προστίθενται στη λίστα **stem_sentence**. Η συνάρτηση επιστρέφει τη λίστα stem_sentence, η οποία περιέχει τα προεπεξεργασμένα και βασικά token που έχουν περάσει από όλα τα βήματα φιλτραρίσματος και καθαρισμού.

Συνολικά, η συνάρτηση tokenize_and_stem χρησιμεύει ως ένα εργαλείο προεπεξεργασίας κειμένου που προετοιμάζει το κείμενο εισαγωγής για περαιτέρω ανάλυση. Επικεντρώνεται στην εξαγωγή λέξεων με νόημα, ενώ απορρίπτει τον θόρυβο,

βελτιώνοντας την ακρίβεια και την αποτελεσματικότητα των επόμενων εργασιών ανάλυσης κειμένου.

Ακολουθεί η συνάρτηση **title_description_filtering**, η οποία εκτελεί μια σειρά βημάτων προεπεξεργασίας σε ένα DataFrame που περιέχει τον τίτλο και την περιγραφή της κάθε είδησης. Η διαδικασία για την περιγραφή είναι ίδια με του τίτλου με μικρές παραλλαγές στον κώδικα και για αυτό παραλείφθηκε από την εικόνα. Αυτή η συνάρτηση δημιουργεί ένα νέο DataFrame με το όνομα **title_vocab_frame**, όπου κάθε σειρά περιέχει σε μία στήλη τη βασική μορφή των λέξεων [KB16] και σε άλλη στήλη την κανονική λέξη η οποία υπέστη επεξεργασία. Αυτό το DataFrame είναι η βασική πηγή που χρησιμοποιήθηκε για την επιλογή των επιπλέον λέξεων τερματισμού που αναφέρθηκε προηγουμένως. Συγκεκριμένα, εκτυπώθηκε σε συνεχόμενες επαναλήψεις του κώδικα, ώστε να αναλυθεί και να εντοπιστούν λέξεις που για τη συγκεκριμένη εργασία είναι περιττές και χωρίς ουσιώδες περιεχόμενο.

```
def title_description_filtering(df):
    # δημιουργία dataframe title_vocab_frame όπου κάθε γραμμή περιέχει την λέξη που έχει υποστεί stemming και την λέξη χωρίς stemming
    title_totalvocab_stemmed = []  # λίστα με όλα τα tokens που έχουν υποστεί stemming
    title_totalvocab_tokenized = []  # λίστα με όλα τα tokens
    title_total_indexes = []
    k = 0

    for i in df['title'].tolist():

        title_allwords_stemmed = tokenize_and_stem(i)  # για κάθε είδηση του dataframe, tokenize και stemming
        title_totalvocab_stemmed.extend(title_allwords_stemmed)  # επέκταση της 'totalvocab_stemmed' λίστας

        title_allwords_tokenized = tokenize_only(i)  # για κάθε είδηση του dataframe, tokenize
        title_totalvocab_tokenized.extend(title_allwords_tokenized)  # επέκταση της 'totalvocab_tokenized' λίστας

        for j in title_allwords_stemmed:
            title_total_indexes.append(k)

        k = k + 1

    title_vocab_frame = pd.DataFrame(
        {'title_stemmed': title_totalvocab_stemmed, 'title_words': title_totalvocab_tokenized},
        index=title_total_indexes)  # δημιουργία του dataframe title_vocab_frame
    #display(title_vocab_frame)
```

Εικόνα 23: Η συνάρτηση title_description_filtering.

Ο σκοπός αυτής της συνάρτησης είναι να δημιουργήσει ένα DataFrame που να συσχετίζει λέξεις που προέρχονται από τη στήλη «τίτλος» με τα αντίστοιχα ευρετήρια τους στο αρχικό DataFrame. Αυτό το βήμα προεπεξεργασίας είναι χρήσιμο για διάφορες εργασίες ανάλυσης κειμένου, όπως η οικοδόμηση ενός λεξιλογίου ή η εκτέλεση περαιτέρω αναλύσεων στις λέξεις που έχουν διαμορφωθεί ως συμβολικές και βασικές λέξεις.

Ακολουθεί η συνάρτηση **pre_process** η οποία πραγματοποιεί προεπεξεργασία κειμένου στους τίτλους των ειδησεογραφικών άρθρων που έχουν συλλεχθεί.

Χρησιμοποιεί τη βιβλιοθήκη spaCy για επεξεργασία φυσικής γλώσσας. Ο πρωταρχικός στόχος αυτής της συνάρτησης είναι η επεξεργασία των τίτλων εισόδου και η επιστροφή μιας λίστας προεπεξεργασμένων, λημματοποιημένων και κωδικοποιημένων εγγράφων.

```
# Προ-επεξεργασία τίτλων με αφαίρεση των stopwords και ληματοποίηση
def pre_process(titles):

    # Προ-επεξεργασία όλων των ειδήσεων
    title_docs = [nlp(x) for x in titles]
    preprocessed_title_docs = []
    lemmatized_tokens = []
    for title_doc in title_docs:
        for token in title_doc:
            if not token.is_stop:
                lemmatized_tokens.append(token.lemma_)
        preprocessed_title_docs.append(" ".join(lemmatized_tokens))
        #άδειασμα της λίστας των ληματοποιημένων tokens μόλις επεξεργαστεί ο επόμενος τίτλος
        del lemmatized_tokens[
            :
            ]

    return preprocessed_title_docs
```

Εικόνα 24: Η συνάρτηση pre_process.

Τα βήματα που πραγματοποιεί η συνάρτηση:

- Η συνάρτηση ξεκινά με την προετοιμασία μιας λίστας που ονομάζεται title_docs για τη διατήρηση των επεξεργασμένων εγγράφων. Κάθε τίτλος υποβάλλεται σε επεξεργασία χρησιμοποιώντας τη γραμμή επεξεργασίας φυσικής γλώσσας του spaCy και κάνει tokenize, αναλύει και προσθέτει ετικέτες στο κείμενο εισόδου.
- Για κάθε έγγραφο τίτλου, η συνάρτηση επαναλαμβάνεται μέσω κάθε token (λέξης) στο έγγραφο. Ελέγχει εάν το token δεν είναι λέξη τερματισμού και τότε λημματοποιείται, δηλαδή μειώνεται στη βασική του μορφή, χρησιμοποιώντας το χαρακτηριστικό λήμμα και προστίθεται σε μια λίστα που ονομάζεται lemmatized_tokens.
- Μετά την επεξεργασία όλων των token σε έναν τίτλο, η συνάρτηση ενώνει τα λημματοποιημένα token σε μια ενιαία συμβολοσειρά και την προσαρτά στη λίστα preprocessed_title_docs. Στη συνέχεια, η λίστα lemmatized_tokens διαγράφεται για να προετοιμαστεί για τον επόμενο τίτλο.
- Τέλος, η συνάρτηση επιστρέφει τη λίστα των προεπεξεργασμένων εγγράφων τίτλου **preprocessed_title_docs**.

Συνολικά, η συνάρτηση pre_process εφαρμόζει τη διοχέτευση NLP του spaCy για την ονοματοποίηση και τη λήμματοποίηση λέξεων στους τίτλους εισαγωγής [DZ11]. Αφαιρεί τις λέξεις τερματισμού και δημιουργεί προεπεξεργασμένα έγγραφα που είναι έτοιμα για περαιτέρω ανάλυση. Αυτό το είδος προεπεξεργασίας βοηθά στη μείωση του θορύβου, στην απλοποίηση των λέξεων και στη βελτίωση της ποιότητας των δεδομένων κειμένου για διάφορες εργασίες ανάλυσης κειμένου.

6.2.3 Επίλογος

Συμπερασματικά, η προεπεξεργασία αποτελεί μια ουσιαστική διαδικασία που μετατρέπει το ακατέργαστο, αδόμητο κείμενο σε μια εκλεπτυσμένη και δομημένη μορφή μέσω τεχνικών όπως η τμηματοποίηση, η μετατροπή στη βασική μορφή των λέξεων, η λημματοποίηση, η αφαίρεση λέξεων τερματισμού και η κανονικοποίηση.

Ωστόσο, η προεπεξεργασία δεν είναι μια διαδικασία που παραμένει ίδια σε κάθε εργασία ανάλυσης κειμένου. Απαιτεί να λαμβάνονται υπόψη τα μοναδικά χαρακτηριστικά των δεδομένων και οι στόχοι της ανάλυσης. Με κατάλληλες προσαρμογές και τροποποιήσεις μπορεί να αυξηθεί περαιτέρω η μετασχηματιστική ισχύς της προεπεξεργασίας, οδηγώντας σε πιο ολοκληρωμένα αποτελέσματα.

6.3 Δημιουργία βάσης δεδομένων και εισαγωγή δεδομένων και μεταδεδομένων ειδησεογραφικών άρθρων

6.3.1 Εισαγωγή

Οι βάσεις δεδομένων είναι το βασικό αποθετήριο δεδομένων για όλες τις εφαρμογές λογισμικού. Για παράδειγμα, κάθε φορά που κάποιος πραγματοποιεί μια αναζήτηση στον ιστό, συνδέεται σε έναν λογαριασμό ή ολοκληρώνει μια συναλλαγή, ένα σύστημα βάσης δεδομένων αποθηκεύει τις πληροφορίες ώστε να είναι δυνατή η πρόσβαση σε αυτές στο μέλλον.

Μια σχεσιακή βάση δεδομένων αποθηκεύει δεδομένα σε ξεχωριστούς πίνακες αντί να τοποθετεί όλα τα δεδομένα σε μια μεγάλη αποθήκη. Η δομή της βάσης δεδομένων είναι οργανωμένη σε φυσικά αρχεία βελτιστοποιημένα για ταχύτητα. Το λογικό μοντέλο δεδομένων, με αντικείμενα, όπως πίνακες δεδομένων, προβολές, σειρές και στήλες, προσφέρει ένα ευέλικτο περιβάλλον προγραμματισμού. Ρυθμίζει κανόνες που διέπουν τις σχέσεις μεταξύ διαφορετικών πεδίων δεδομένων, όπως ένα προς ένα, ένα προς πολλά, μοναδικά, υποχρεωτικά ή προαιρετικά και "δείκτες" μεταξύ

διαφορετικών πινάκων. Η βάση δεδομένων επιβάλλει αυτούς τους κανόνες έτσι ώστε με μια καλά σχεδιασμένη βάση δεδομένων η χρήση μιας εφαρμογής να μην βλέπει ποτέ δεδομένα που είναι ασυνεπή, διπλότυπα, ορφανά, παλιά ή λείπουν.

Το τμήμα "SQL" του "MySQL" σημαίνει "Structured Query Language". Η SQL είναι η πιο κοινή τυποποιημένη γλώσσα που χρησιμοποιείται για την πρόσβαση σε βάσεις δεδομένων. Ανάλογα με το περιβάλλον προγραμματισμού, μπορεί να γίνει απευθείας εισαγωγή SQL (για παράδειγμα, για την δημιουργία αναφορών), να γίνει ενσωμάτωση δηλώσεων SQL σε κώδικα γραμμένο σε άλλη γλώσσα ή να γίνει χρήση ενός API συγκεκριμένης γλώσσας που κρύβει τη σύνταξη SQL [Ora23b], [B23c].

6.3.2 MySQL και Workbench

Η MySQL είναι η κατεξοχήν βάση δεδομένων ανοικτού κώδικα παγκοσμίως, γεγονός που σημαίνει ότι το λογισμικό είναι προσβάσιμο από οποιονδήποτε να το χρησιμοποιήσει και να το προσαρμόσει ανάλογα με τις ανάγκες του. Οι χρήστες έχουν την ελευθερία να αποκτήσουν τη MySQL από το διαδίκτυο χωρίς κανένα κόστος. Η χρήση του λογισμικού διέπεται από τη Γενική Άδεια Δημόσιας Χρήσης GNU (GNU General Public License, GPL), η οποία υπαγορεύει τις επιτρεπτές ενέργειες και τους περιορισμούς που σχετίζονται με το λογισμικό υπό διάφορες συνθήκες.

Όπως αναφέρει η DB-Engines, η MySQL κατέχει τη δεύτερη θέση μεταξύ των πιο διαδεδομένων βάσεων δεδομένων, με την Oracle Database να έχει την πρωτιά. Η MySQL χρησιμεύει ως ραχοκοκαλιά για πολυάριθμες ευρέως χρησιμοποιούμενες εφαρμογές, όπως το Facebook, το Twitter, το Netflix, η Uber, η Airbnb, το Spotify και η Booking.com. Δεδομένου του χαρακτήρα της ως ανοικτού κώδικα, η MySQL προσφέρει μια σειρά από χαρακτηριστικά που έχουν εξελιχθεί κατά τη διάρκεια ενός τετάρτου αιώνα (25 χρόνια) στενής συνεργασίας με τους χρήστες. Κατά συνέπεια, είναι πολύ πιθανό η MySQL να υποστηρίζει την εν λόγω εφαρμογή ή γλώσσα προγραμματισμού. [Ora23b].

Η MySQL αποτελεί μια ιδιαίτερα διαδεδομένη επιλογή βάσης δεδομένων, που προτιμάται από οργανισμούς για την τροφοδοσία ζωτικών επιχειρηματικών εφαρμογών λόγω της εξαιρετικής αξιοπιστίας της. Προσαρμόζεται απρόσκοπτα για να ανταποκρίνεται στις απαιτήσεις των εφαρμογών με μεγάλη κυκλοφορία. Το εγγενές πλαίσιο αντιγραφής της δίνει τη δυνατότητα σε οργανισμούς, όπως το Facebook, να επεκτείνουν τις εφαρμογές τους για να φιλοξενήσουν τεράστιες βάσεις χρηστών, με τη

MySQL να παρέχει μια ολοκληρωμένη “σουίτα” πλήρως ενσωματωμένων εργαλείων αντιγραφής για τη διασφάλιση ισχυρών λύσεων υψηλής διαθεσιμότητας και ανάκαμψης από καταστροφές.

Οι πελάτες μπορούν να επιτύχουν την τήρηση των συμφωνιών επιπέδου εξυπηρέτησης και την ικανοποίηση των αναγκών των κρίσιμων εφαρμογών. Η ασφάλεια δεδομένων συνεπάγεται τη διασφάλιση και την τήρηση των βιομηχανικών και κυβερνητικών εντολών, όπως το Πρότυπο Ασφάλειας Δεδομένων της Βιομηχανίας Καρτών Πληρωμών, ο Νόμος περί Φορητότητας και Ευθύνης των Ασφαλιστικών Οργανισμών Υγείας και οι Οδηγίες Τεχνικής Εφαρμογής Ασφάλειας της Υπηρεσίας Πληροφοριακών Συστημάτων Άμυνας. Η MySQL Enterprise Edition προσφέρει μια σειρά προηγμένων λειτουργιών ασφαλείας, όπως έλεγχο ταυτότητας και εξουσιοδότηση, έλεγχο και τείχος προστασίας βάσεων δεδομένων.

Η συμμετοχή της MySQL στο έργο αυτό οφείλεται στην ταχύτητα, την αξιοπιστία, την επεκτασιμότητα και τη φιλικότητα προς το χρήστη. Η MySQL σχεδιάστηκε αρχικά για την αποτελεσματική διαχείριση μεγάλων βάσεων δεδομένων και έχει αποδείξει την απόδοσή της σε απαιτητικά περιβάλλοντα παραγωγής με την πάροδο των ετών. Παρά τη συνεχή ανάπτυξή της, η MySQL διαθέτει ένα σημαντικό και πρακτικό σύνολο χαρακτηριστικών. Με την εξαιρετική συνδεσιμότητα, την ταχύτητα και την ασφάλειά της, η MySQL είναι μια εξαιρετικά κατάλληλη επιλογή για πρόσβαση σε βάσεις δεδομένων μέσω διαδικτύου. [Ora23b].

Είναι σημαντικό να τονιστεί ότι η χρήση της MySQL, μαζί με την οπτικοποίησή της, απαιτεί τη χρήση του MySQL Workbench. Το MySQL Workbench χρησιμεύει ως μια ενοποιημένη οπτική εφαρμογή που απευθύνεται σε αρχιτέκτονες βάσεων δεδομένων, προγραμματιστές και DBAs (Data Base Administrator (διαχειριστές βάσεων δεδομένων)). Αυτό το εργαλείο περιλαμβάνει χαρακτηριστικά για τη μοντελοποίηση δεδομένων, την ανάπτυξη SQL και ολοκληρωμένες εργασίες διαχείρισης, συμπεριλαμβανομένων των ρυθμίσεων του διακομιστή, της διαχείρισης χρηστών, των αντιγράφων ασφαλείας και πρόσθετων λειτουργιών.

Το MySQL Workbench δίνει τη δυνατότητα σε έναν διαχειριστή βάσεων δεδομένων (DBA), έναν προγραμματιστή ή έναν αρχιτέκτονα δεδομένων να σχεδιάζει, να δημιουργεί και να επιβλέπει βάσεις δεδομένων με οπτικό τρόπο. Προσφέρει ένα ολοκληρωμένο σύνολο εργαλείων για τους μοντελοποιητές δεδομένων για την

κατασκευή περίπλοκων μοντέλων ER, τη διεξαγωγή τόσο εμπρόσθιας όσο και αντίστροφης μηχανικής και απλοποιεί απαιτητικές εργασίες διαχείρισης αλλαγών και τεκμηρίωσης που παραδοσιακά απαιτούν σημαντικό χρόνο και προσπάθεια. Επιπλέον, το MySQL Workbench παρέχει οπτικά μέσα για τη δημιουργία, εκτέλεση και βελτίωση ερωτημάτων SQL. Ο SQL Editor εντός του εργαλείου περιλαμβάνει χαρακτηριστικά, όπως η έγχρωμη επισήμανση συντακτικού, η αυτόματη συμπλήρωση, η δυνατότητα επαναχρησιμοποίησης αποσπασμάτων κώδικα SQL και η διατήρηση ιστορικού των εκτελέσεων ερωτημάτων SQL.

Επιπρόσθετα, το MySQL Workbench προσφέρει μια οπτική διεπαφή που απλοποιεί τη διαχείριση των περιβαλλόντων MySQL και ενισχύει τη διαφάνεια της βάσης δεδομένων. Οι προγραμματιστές και οι DBA μπορούν να χρησιμοποιούν αυτά τα οπτικά εργαλεία για να ρυθμίζουν διακομιστές, να διαχειρίζονται λογαριασμούς χρηστών, να εκτελούν εργασίες δημιουργίας αντιγράφων ασφαλείας και αποκατάστασης, να εξετάζουν πληροφορίες ελέγχου και να παρακολουθούν την κατάσταση των βάσεων δεδομένων με μεγαλύτερη ευκολία. [Ora23c].

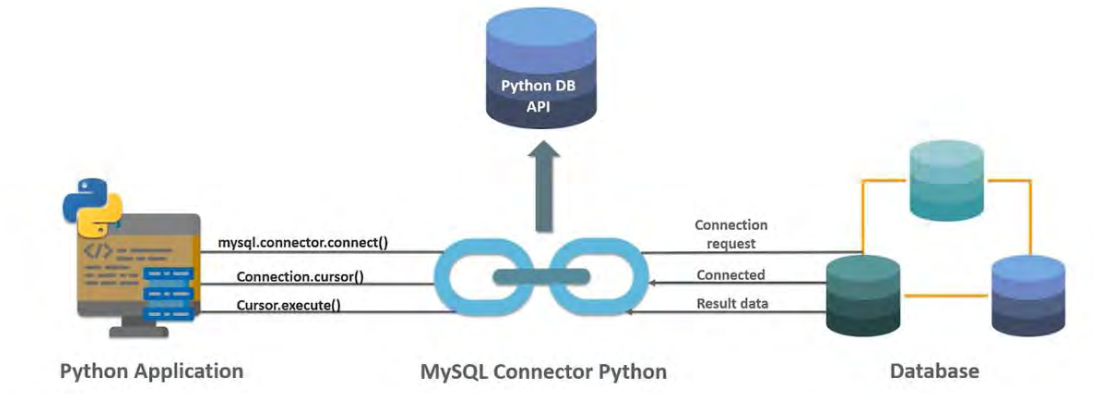
6.3.3 Διαδικασία δημιουργίας και σύνδεσης της βάσης με τον κώδικα

Για την πραγματοποίηση της καταγραφής και αποθήκευσης των ειδήσεων, απαραίτητη κρίνεται η δημιουργία μια βάσης δεδομένων καθώς τα γεγονότα που αντλούνται καθημερινά από την συγκεκριμένη πτυχιακή εργασία είναι χιλιάδες. Αυτό, έχει ως αποτέλεσμα, την καλύτερη οργάνωση των δεδομένων καθώς και την πιο εύκολη διαχείρισή τους χρησιμοποιώντας την δομή των πινάκων, μια δυνατότητα που προσφέρεται από την MySQL.

Πρώτα από όλα θα πρέπει να αναλυθεί η λογική πίσω από την δημιουργία της σύνδεσης του κώδικα σε Python με την βάση δεδομένων MySQL. Για την καλύτερη κατανόηση των διαδικασιών, χρησιμοποιήθηκε η παρακάτω εικόνα (Εικόνα 25) ώστε να φαίνεται και οπτικά πως λειτουργεί το μέρος της σύνδεσης. Αναλυτικότερα, αριστερά φαίνεται η βάση δεδομένων, δεξιά η εφαρμογή σε Python και στην μέση ο συνδέτης της MySQL με την Python. Ο συνδέτης της MySQL με την Python λειτουργεί ως γέφυρα μεταξύ της front-end εφαρμογής σε Python και του back-end server της βάσης δεδομένων.

Συγκεκριμένα, αυτό που συμβαίνει είναι πως η front-end εφαρμογή σε Python θα στείλει ένα αίτημα σύνδεσης (mysql.connector.connect()) στον συνδέτη και μετά ο

συνδέτης θα στείλει το ίδιο αίτημα σύνδεσης (Connection request) στη βάση δεδομένων. Στη συνέχεια η βάση θα αποδεχτεί το αίτημα και θα στείλει ένα μήνυμα σύνδεσης (Connected) στον συνδέτη, όπου εκείνος με την σειρά του θα στείλει ένα αίτημα σύνδεσης δρομέα (Connection.cursor()). Οπότε η λογική της σύνδεσης ως μέθοδο και ο δρομέα ως αντικείμενο θα οδηγήσει στην επικοινωνία με ολόκληρη τη βάση δεδομένων MySQL αλλά και θα επιτρέψει την δημιουργία της (δικής μας) βάσης δεδομένων. Τέλος, η εφαρμογή σε Python θα εκτελέσει μια συγκεκριμένη δήλωση συνέχειας ή ένα συγκεκριμένο ερώτημα συνέχειας (Cursor.execute()) στον συνδέτη και ο server της βάσης δεδομένων θα ανακτήσει τα δεδομένα αποτελέσματος (Result data).



Εικόνα 25: Σύνδεση της Python με την Βάση Δεδομένων MySQL. Πηγή: https://www.youtube.com/watch?v=g60QghtJmjY

Στη συνέχεια θα ήταν αξιοσημείωτο να αναφερθούν οι λειτουργίες που μπορούν να εκτελεστούν μεταξύ της Python και της MySQL. Οι λειτουργίες αυτές ονομάζονται C.R.U.D. (Create, Read, Update, Delete) δηλαδή μπορούν να πραγματοποιηθεί η δημιουργία, η ανάγνωση, η ενημέρωση και η διαγραφή της βάσης δεδομένων μέσω της Python.

Προκειμένου, λοιπόν, να γίνει πιο εύκολη και κατανοητή η διαχείριση των δεδομένων μας, είναι σημαντική η δημιουργία μιας βάσης δεδομένων. Η Python είναι μια ευρέως χρησιμοποιούμενη γλώσσα προγραμματισμού στην επιστημονική κοινότητα, όμως η ανάπτυξη μιας εφαρμογής είναι αδύνατη χωρίς να μην είναι γνωστό το μέρος που αποθηκεύονται τα δεδομένα. Σε αυτό το σημείο, είναι απαραίτητη, η διαχείριση ενός συστήματος βάσεων δεδομένων και ένας από τους πιο γνωστούς DBMS (DataBase Management System) που χρησιμοποιούνται στην βιομηχανία είναι ο MySQL DB Server.

6.3.3.1 Δημιουργία και σύνδεση της βάσης δεδομένων με τον κώδικα

Ακολουθεί το κομμάτι του κώδικα που χρησιμοποιείται για την δημιουργία της βάσης δεδομένων καθώς και της σύνδεσης του κώδικα με την βάση δεδομένων της συγκεκριμένης εργασίας.

```
mydb = mysql.connector.connect(host="localhost", user="root", passwd="rozmarispiridoula2022!",
                               database="greekeventdb")
if (mydb):
    print("Connection successful")
else:
    print("Connection unsuccessful")

mycursor = mydb.cursor(buffered=True)
```

Εικόνα 26: Δημιουργία και σύνδεση της βάσης δεδομένων με τον κώδικα.

Σε αυτό το μπλοκ κώδικα, δημιουργείται μια σύνδεση με τη βάση δεδομένων MySQL. Η σημασία κάθε παραμέτρου φαίνεται παρακάτω:

- **¨host¨**: Το όνομα κεντρικού υπολογιστή ή η διεύθυνση IP του διακομιστή MySQL. Σε αυτήν την περίπτωση, ορίζεται σε "localhost", που συνήθως αναφέρεται στον τοπικό υπολογιστή.
- **¨user¨**: Το όνομα χρήστη που χρησιμοποιείται για τη σύνδεση στον διακομιστή MySQL. Σε αυτή την περίπτωση, είναι "root".
- **¨passwd¨**: Ο κωδικός πρόσβασης που σχετίζεται με το συγκεκριμένο όνομα χρήστη. Σε αυτή την περίπτωση, είναι "rozmarispiridoula2022!"
- **¨database¨**: Το όνομα της βάσης δεδομένων στην οποία θα γίνει η σύνδεση. Σε αυτή την περίπτωση, είναι "greekeventdb".

Αφού γίνει η προσπάθεια δημιουργία της σύνδεσης, ο κωδικός ελέγχει εάν η σύνδεση ήταν επιτυχής. Η συνθήκη if ελέγχει εάν το **mydb** έχει μια τιμή (που υποδεικνύει μια επιτυχημένη σύνδεση) ή είναι None (υποδηλώνει ανεπιτυχή σύνδεση). Ανάλογα με το αποτέλεσμα, εκτυπώνει είτε "Connection successful" είτε "Connection unsuccessful" στην κονσόλα.

mycursor = mydb.cursor(buffered=True): Μόλις δημιουργηθεί η σύνδεση, δημιουργείται ένας δρομέας χρησιμοποιώντας τη μέθοδο cursor() του αντικειμένου σύνδεσης της βάσης δεδομένων (mydb). Ο δρομέας είναι ένα αντικείμενο βάσης δεδομένων που χρησιμοποιείται για την αλληλεπίδραση με τη βάση δεδομένων και την εκτέλεση ερωτημάτων SQL. Σε αυτήν την περίπτωση, η παράμετρος προσωρινής αποθήκευσης ορίζεται σε True, πράγμα που σημαίνει ότι ο δρομέας θα ανακτήσει και

θα αποθηκεύσει σειρές προσωρινής αποθήκευσης για τα ερωτήματα που εκτελούνται, επιτρέποντοντας στον χρήστη την πλοήγηση στο σύνολο αποτελεσμάτων.

Συνολικά, αυτός ο κώδικας δημιουργεί μια σύνδεση με τη βάση δεδομένων MySQL "greekeventdb", ελέγχει εάν η σύνδεση ήταν επιτυχής και δημιουργεί ένα αντικείμενο δρομέα που μπορεί να χρησιμοποιηθεί για αλληλεπίδραση με τη βάση δεδομένων εκτελώντας ερωτήματα SQL.

6.3.3.2 Δημιουργία μιας μηχανής (engine) βάσης δεδομένων

Στη συνέχεια, αναγκαία είναι η δημιουργία μιας μηχανής (engine) βάσης δεδομένων για σύνδεση στη βάση δεδομένων MySQL "greekeventdb". Αυτό, επιτυγχάνεται με την παρακάτω γραμμή κώδικα.

```
engine = create_engine('mysql+mysqlconnector://root:rozmarispiridoula2022!@localhost:3306/greekeventdb', echo=False)
```

Εικόνα 27: Δημιουργία μιας μηχανής (engine) βάσης δεδομένων.

Η συνάρτηση "**create_engine**" χρησιμοποιείται για τη δημιουργία μιας μηχανής βάσης δεδομένων. Αυτή η μηχανή είναι ένα κρίσιμο στοιχείο που διαχειρίζεται τις συνδέσεις και τις αλληλεπιδράσεις της βάσης δεδομένων. Η ανάλυση των ορισμάτων στην κλήση της συνάρτησης είναι η εξής:

'mysql+mysqlconnector://root:rozmarispiridoula2022!@localhost:3306/greekeventdb': Αυτό το τμήμα καθορίζει τη διεύθυνση URL σύνδεσης για τη βάση δεδομένων MySQL. Ακολουθεί μια μορφή που είναι γενικά τυποποιημένη στο SQLAlchemy:

- **"mysql+mysqlconnector"**: Υποδεικνύει τη διάλεκτο (τύπο) της βάσης δεδομένων με την οποία πρέπει να συνδεθεί ο χρήστης, η οποία είναι η MySQL σε αυτήν την περίπτωση. Το SQLAlchemy υποστηρίζει διάφορες διαλέκτους βάσης δεδομένων.
- **"root:rozmarispiridoula2022!"**: Αυτό είναι το όνομα χρήστη και ο κωδικός πρόσβασης για τη σύνδεση της βάσης δεδομένων, διαχωρισμένα με άνω και κάτω τελεία. Το όνομα χρήστη είναι **"root"** και ο κωδικός πρόσβασης είναι **"rozmarispiridoula2022!"**
- **"localhost:3306"**: Καθορίζει τον κεντρικό υπολογιστή και τη θύρα όπου εκτελείται η βάση δεδομένων MySQL. Το **"localhost"** υποδεικνύει ότι η βάση δεδομένων φιλοξενείται στον ίδιο υπολογιστή όπου εκτελείται ο κώδικας και ότι το 3306 είναι η προεπιλεγμένη θύρα για τη MySQL.

- **“/greekeventdb”**: Αυτό το τμήμα της διεύθυνσης URL υποδεικνύει τη συγκεκριμένη βάση δεδομένων εντός του διακομιστή MySQL στον οποίο είναι επιθυμητή η σύνδεση.
- **“echo=False”**: Η παράμετρος echo ελέγχει εάν οι εντολές SQL που εκτελούνται από την μηχανή επαναλαμβάνονται (εκτυπώνονται) στην κονσόλα. Η ρύθμιση του σε False απενεργοποιεί αυτήν την ηχώ.

Συνοπτικά, αυτός ο κώδικας χρησιμοποιεί τη συνάρτηση “**create_engine”** του SQLAlchemy για να δημιουργήσει μια μηχανή βάσης δεδομένων που συνδέεται με μια βάση δεδομένων MySQL με το όνομα "greekeventdb" στον τοπικό υπολογιστή χρησιμοποιώντας το παρεχόμενο όνομα χρήστη ("root") και κωδικό πρόσβασης ("rozmarispiridoula2022!"). Η μηχανή μπορεί να χρησιμοποιηθεί για την εκτέλεση ερωτημάτων SQL και την αλληλεπίδραση με τη βάση δεδομένων.

6.3.3.3 Δημιουργία πινάκων στη βάση μέσω του κώδικα σε Python

Η δημιουργία πινάκων είναι απαραίτητη για την αποθήκευση των δεδομένων μέσα στη βάση με οργανωμένο τρόπο. Συγκεκριμένα, δημιουργήθηκαν 11 διαφορετικοί πίνακες, καθώς αναλύουμε γεγονότα που ανήκουν σε 11 μεγάλες κατηγορίες οι οποίες αναλύονται στο επόμενο κεφάλαιο. Η δημιουργία των πινάκων μέσα στη βάση γίνεται με εύκολο και κατανοητό τρόπο ορίζοντας κατά την δημιουργία του και τις στήλες που θα έχει ο πίνακας.

Παρακάτω εμφανίζεται ο κώδικας για την δημιουργία του πίνακα της κατηγορίας “crime_accident”, που ουσιαστικά είναι ένα ερώτημα SQL που εκτελείται χρησιμοποιώντας τη μέθοδο **execute** σε ένα αντικείμενο δρομέα. Το ερώτημα δημιουργεί έναν νέο πίνακα με το όνομα **crime_accident** στη βάση δεδομένων.

```
mycursor.execute("Create table crime_accident(id INT AUTO_INCREMENT PRIMARY KEY,
                 title text, category varchar(200), description mediumtext, link text,
                 content longtext, pubdate varchar(100), sublevel varchar(100), original_news_id INT)")
```

Εικόνα 28: Δημιουργία του πίνακα crime_accident στη βάση δεδομένων. Για λόγους εμφάνισης της εικόνας, το συγκεκριμένο query της MySQL δεν βρίσκεται σε μία γραμμή, εξού η κόκκινη υπογράμμιση.

Αυτό το ερώτημα SQL δημιουργεί έναν πίνακα με το όνομα **“crime_accident”** με τις ακόλουθες στήλες:

- **“id”:** Μια ακέραια στήλη που χρησιμεύει ως το πρωτεύον κλειδί του πίνακα. Η λέξη-κλειδί **“AUTO_INCREMENT”** υποδεικνύει ότι αυτή η στήλη θα

δημιουργήσει αυτόματα μοναδικές τιμές καθώς προστίθενται νέες σειρές. Αυτή η στήλη θα προσδιορίζει μοναδικά κάθε σειρά στον πίνακα.

- **“title”:** Μια στήλη τύπου **“TEXT”** (συμβολοσειρά μεταβλητού μήκους) για την αποθήκευση του τίτλου ενός εγκλήματος ή ενός ατυχήματος.
- **“category”:** Μια στήλη τύπου **“VARCHAR(200)”** (συμβολοσειρά μεταβλητού μήκους με μέγιστο μήκος 200 χαρακτήρες) για την αποθήκευση της κατηγορίας ή του τύπου του εγκλήματος ή του ατυχήματος.
- **“description”:** Μια στήλη τύπου **“MEDUIUMTEXT”** για την αποθήκευση μιας πιο λεπτομερούς περιγραφής του εγκλήματος ή του ατυχήματος.
- **“link”:** Μια στήλη τύπου **“TEXT”** για την αποθήκευση ενός συνδέσμου που σχετίζεται με το έγκλημα ή το ατύχημα.
- **“content”:** Μια στήλη τύπου **“LONGTEXT”** για αποθήκευση δυνητικά μεγάλου περιεχομένου που σχετίζεται με το έγκλημα ή το ατύχημα.
- **“pubdate”:** Μια στήλη τύπου **“VARCHAR(100)”** για την αποθήκευση της ημερομηνίας δημοσίευσης του εγκλήματος ή του ατυχήματος.
- **“sublevel”:** Μια στήλη τύπου **“VARCHAR(100)”** για την αποθήκευση ενός αναγνωριστικού sublevel που σχετίζεται με την κατηγορία που κατατάσσεται το γεγονός.
- **“original_news_id”:** Μια ακέραια στήλη για την αποθήκευση του αναγνωριστικού (id) του αρχικού άρθρου ειδήσεων ή της πηγής που σχετίζεται με το έγκλημα ή το ατύχημα.

Στην παρακάτω εικόνα απεικονίζεται η μορφή του πίνακα crime_accident μέσα στον πίνακα.

id	title	category	description	link	content	pubdate	sublevel	original_news_id
27	Ουγκάντα: Ενάφρινε η βουλή το νομοσχέδιο κα...	Κόσμος	Πριν από λίγες ημέρες ο πρόεδρος Μουσεβεν, π...	https://www.pagenews.gr/2023/03/22/kosmos...	Πριν από λίγες ημέρες ο πρόεδρος Μουσεβενι, π...	22-03-2023 09:34:09	1	0
28	Επίθεση-σοκ από μαχαιροβγάλτες έξω από το Ε...	ΕΛΛΑΔΑ	Αιματηρή επίθεση από κουκουλοφόρους σε νεα...	https://www.dimokratia.gr/ellada/558628/epith...	Αιματηρή επίθεση από κουκουλοφόρους σε νεα...	23-03-2023 11:09:07	1	0
29	Η δημοσιογραφία στα μέτρα της εξουσίας	Η ΘΕΣΗ ΜΑΣ	Η τηλεοπτική συνέντευξη του Κυριάκου Μητσοτ...	https://www.dimokratia.gr/i-thesi-mas/558597/i...	Η τηλεοπτική συνέντευξη του Κυριάκου Μητσοτ...	23-03-2023 09:29:37	1	0
30	Τζόκσον Πόλοκ: Πίνακάς του βρέθηκε στη Βουλ...	Πολιτισμός	Τζόκσον Πόλοκ: Τα άγνωστα μέχρι σήμερα εργ...	https://www.pagenews.gr/2023/03/23/politism...	Τζόκσον Πόλοκ: Τα άγνωστα μέχρι σήμερα εργ...	23-03-2023 11:25:10	1	0
31	Ξανά στη φυλακή! Νέα δίκη για τους επτά του ...	ΕΛΛΑΔΑ	Ο Άρειος Πάγος αναίρεσε την απόφαση του Εφε...	https://www.dimokratia.gr/ellada/558514/xana...	Ο Άρειος Πάγος αναίρεσε την απόφαση του Εφε...	21-03-2023 13:12:06	1	0
32	Άννα Μισέλ Ασημακοπούλου στα pagenews.gr:...	Opinion Leader	Η Ευρωβουλευτής της Νέας Δημοκρατίας κ. Αν...	https://www.pagenews.gr/2023/03/22/opinion-...	Η Ευρωβουλευτής της Νέας Δημοκρατίας κ. Αν...	22-03-2023 18:55:08	1.2.2	0
33	Επίθεση της αναρχικής ομάδας του «Ρουβίκωνα:...	Πολιτική	Τα μέλη της αναρχικής ομάδας το βράδυ της Τρί...	https://www.pagenews.gr/2023/03/22/politiki/...	Τα μέλη της αναρχικής ομάδας το βράδυ της Τρί...	22-03-2023 08:30:48	1	0
34	Αλλοδαπός εξέδιδε και βιντεοσκοπούσε 14χρονη	ΕΛΛΑΔΑ	Ο 28χρονος Ταχίρ, που έκλεινε τα ραντεβού μέ...	https://www.dimokratia.gr/ellada/558554/allod...	Ο 28χρονος Ταχίρ, που έκλεινε τα ραντεβού μέ...	22-03-2023 10:45:47	1.1.1.2	0
35	Κυριάκος: «Θα φύγει κάποια στιγμή ο Λουτσέσκ...	Αθλητισμός	Ο υπεύθυνος επικοινωνίας της ΠΑΕ ΠΑΟΚ, Κυρι...	https://www.pagenews.gr/2023/03/22/athlitis...	Ο υπεύθυνος επικοινωνίας της ΠΑΕ ΠΑΟΚ, Κυρι...	22-03-2023 10:36:50	1	0
38	Πυρά κατά Μητσοτάκη για Τέμπη και εκλογές	ΠΟΛΙΤΙΚΗ	Η αντιπολίτευση τον κατηγόρησε για τις ευθύνε...	https://www.dimokratia.gr/politiki/558613/pyra...	Η αντιπολίτευση τον κατηγόρησε για τις ευθύνε...	23-03-2023 10:11:35	1	0
39	Η UEFA ξρανά έρευνα για την Μπαρτσελόνα και...	Αθλητισμός	Η UEFA αποφάσισε να ξρανήσει και η ίδια έρευν...	https://www.pagenews.gr/2023/03/23/athlitis...	Η UEFA αποφάσισε να ξρανήσει και η ίδια έρευν...	23-03-2023 15:44:30	1	0
40	Κοινωνικό «γυφάιστεα» η Γαλλία: Άγριες συγκρ...	Κόσμος	Τραυματίες 28 αστυνομικοί και διαδηλωτές εκ τ...	https://www.pagenews.gr/2023/03/26/kosmos...	Τραυματίες 28 αστυνομικοί και διαδηλωτές εκ τ...	26-03-2023 02:44:26	1	0
47	Αποφάσισε ο Κλοπ για Τσιμίκα – “Φλερτ” από κα...	Αθλητισμός	Ο Έλληνας όσος δεν παίζει όσο θα ήθελε στους ...	https://www.pagenews.gr/2023/03/26/athlitis...	Ο Έλληνας όσος δεν παίζει όσο θα ήθελε στους ...	26-03-2023 12:59:38	1.1.2.3	0
48	Φλόριντα: Γονείς Κατηγόρησαν την δασκάλα γι...	Κόσμος	Στην πραγματικότητα, τα παιδιά είχαν δει φωτο...	https://www.pagenews.gr/2023/03/26/kosmos...	Στην πραγματικότητα, τα παιδιά είχαν δει φωτο...	26-03-2023 11:28:49	1.1.5	0

Εικόνα 29: Δομή του πίνακα κατηγορίας crime_accident μέσα στη βάση δεδομένων.

Αυτή η δομή πίνακα έχει σχεδιαστεί για να αποθηκεύει πληροφορίες σχετικά με συμβάντα εγκλημάτων και ατυχημάτων. Κάθε γραμμή στον πίνακα θα αντιπροσωπεύει ένα συγκεκριμένο συμβάν εγκλήματος ή ατυχήματος και οι στήλες αποθηκεύουν διάφορα χαρακτηριστικά αυτών των συμβάντων, όπως τίτλο, κατηγορία, περιγραφή κ.λπ. Η στήλη id χρησιμεύει ως πρωτεύον κλειδί, διασφαλίζοντας τη μοναδικότητα για κάθε σειρά.

Αυτή η εντολή χρησιμοποιείται και για τις υπόλοιπες 10 κατηγορίες γεγονότων και η μόνη αλλαγή που γίνεται είναι στην κλήση του πίνακα σε αυτή την εντολή.

6.3.3.4 Προετοιμασία των γεγονότων πριν την εισαγωγή τους στη βάση δεδομένων

Στη συγκεκριμένη ενότητα αναφέρεται ένα πολύ σημαντικό κομμάτι της επεξεργασίας των δεδομένων των ειδήσεων προκειμένου να επιτευχθεί με επιτυχία η εισαγωγή τους στη βάση δεδομένων.

6.3.3.4.1 Αφαίρεση του αναγνωριστικού (id) από κάθε γεγονός

Ένα αξιοσημείωτο βήμα της επεξεργασίας των γεγονότων, πριν γίνει η εισαγωγή τους στη βάση, αποτελεί η αφαίρεση του αναγνωριστικού (id) που έχει η κάθε είδηση μέσα στο dataframe. Η διαδικασία αυτή εκτελείται διότι η στήλη id μέσα στην βάση δεδομένων έχει καθοριστεί να αυξάνεται αυτόματα, με αποτέλεσμα να μην είναι δυνατή η αποθήκευση ενός γεγονότος στη βάση που θα έχει το ίδιο id με ένα άλλο γεγονός που υπάρχει ήδη μέσα στη βάση. Βέβαια, αυτό συμβαίνει επειδή ανά μία ώρα που γίνεται η συλλογή των ειδήσεων, τα indexes μηδενίζονται και ξεκινά η αρίθμηση από την αρχή. Έτσι χρησιμοποιείται η εντολή **total_crime_accident_df = total_crime_accident_df.drop('id', axis=1)** αφαιρώντας την στήλη id από το dataframe της κατηγορίας των εγκλημάτων-ατυχημάτων ώστε τα δεδομένα να είναι έτοιμα για εισαγωγή στη βάση.

```
total_crime_accident_df = total_crime_accident_df.drop('id', axis=1)
```

Εικόνα 30: Απεικόνιση της εντολής αφαίρεσης του id στον κώδικα

Φυσικά, αυτή η εντολή χρησιμοποιείται και για τις υπόλοιπες 10 κατηγορίες γεγονότων και η μόνη αλλαγή που γίνεται είναι στα dataframe που αλλάζουν από κατηγορία σε κατηγορία.

6.3.3.5 Εισαγωγή των δεδομένων στη βάση δεδομένων

Στο στάδιο αυτό τα δεδομένα είναι έτοιμα να εισαχθούν και να αποθηκευτούν στη βάση δεδομένων χρησιμοποιώντας την μέθοδο **to_sql**. Η μέθοδος αυτή γράφει τα δεδομένα από ένα dataframe (total_crime_accident_df) σε έναν πίνακα βάσης δεδομένων MySQL που ονομάζεται **'crime_accident'**. Ο κώδικας που κάνει αυτή την λειτουργία φαίνεται παρακάτω.

```
total_crime_accident_df.to_sql(name='crime_accident', con=engine, if_exists='append', index=False)
```

Εικόνα 31: Χρήση της μεθόδου to_sql για την εισαγωγή των δεδομένων στη βάση

Η γραμμή αυτή κώδικα, ουσιαστικά, δείχνει την μετατροπή του dataframe σε SQL και αναλύεται ως εξής:

- **“name='crime_accident'”**: Καθορίζει το όνομα του πίνακα βάσης δεδομένων προορισμού. Σε αυτήν την περίπτωση, πρόκειται για «έγκλημα_ατύχημα».
- **“con=engine”**: Καθορίζει τη σύνδεση βάσης δεδομένων που θα χρησιμοποιηθεί για τη λειτουργία. Η μεταβλητή μηχανή πρέπει να είναι μια μηχανή βάσης δεδομένων SQLAlchemy που αντιπροσωπεύει τη σύνδεση με τη βάση δεδομένων MySQL.
- **“if_exists='append'”**: Καθορίζει τον τρόπο χειρισμού των υπαρχόντων δεδομένων στον πίνακα προορισμού. Η τιμή **"append"** υποδεικνύει ότι τα δεδομένα του dataframe πρέπει να προστεθούν στα υπάρχοντα δεδομένα στον πίνακα. Εάν ο πίνακας δεν υπάρχει, θα δημιουργηθεί.
- **“index=False”**: Υποδεικνύει ότι το ευρετήριο του dataframe δεν πρέπει να περιλαμβάνεται ως ξεχωριστή στήλη στον πίνακα της βάσης δεδομένων. Εάν το ορίσετε σε True, το ευρετήριο του dataframe θα γίνει μια πρόσθετη στήλη στον πίνακα της βάσης δεδομένων.

Συνοπτικά, η παρεχόμενη γραμμή κώδικα παίρνει τα δεδομένα από το dataframe **total_crime_accident_df** και τα γράφει σε έναν πίνακα βάσης δεδομένων MySQL που ονομάζεται **"crime_accident"**. Η μέθοδος **to_sql** επιτρέπει εύκολα την εξαγωγή δεδομένων dataframe σε έναν πίνακα βάσης δεδομένων, ενώ γίνεται έλεγχος διαφόρων πτυχών της διαδικασίας εισαγωγής.

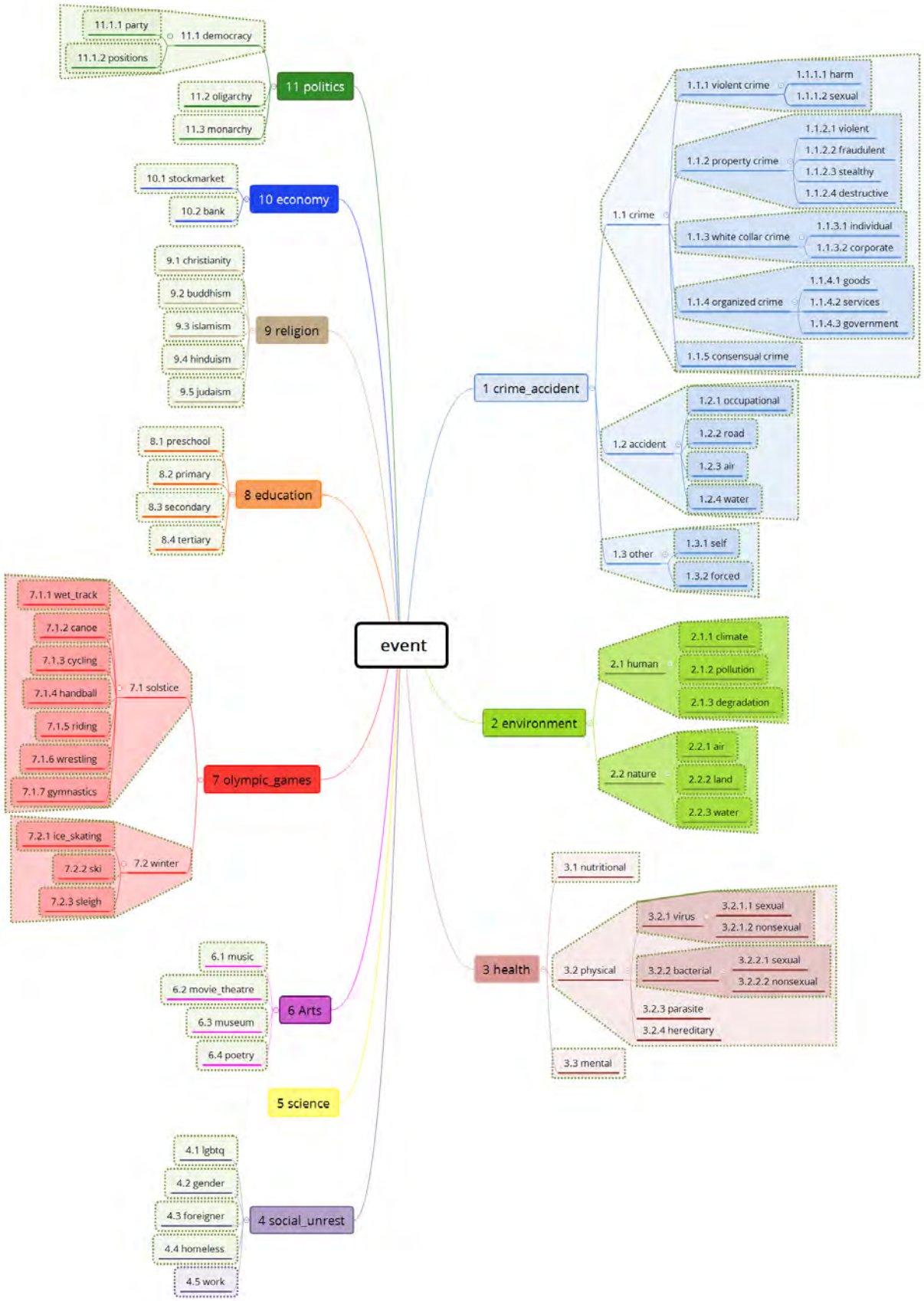
6.3.4 Επίλογος

Η MySQL χρησιμεύει ως ακρογωνιαίος λίθος στη συνεχώς εξελισσόμενη σφαίρα της διαχείρισης δεδομένων. Η εκτεταμένη χρήση της σε ποικίλες εφαρμογές και η υποστήριξή της από ηγέτες του κλάδου υπογραμμίζουν την προσαρμοστικότητα και την αξιοπιστία της. Η δημοτικότητα της MySQL μπορεί να αποδοθεί εν μέρει στην ευελιξία και τη φιλική προς το χρήστη φύση της.

Με βάση την ανασκόπηση του κώδικα που πραγματοποιήθηκε για την παρούσα εργασία, η σύνδεση με τη βάση δεδομένων MySQL δημιουργήθηκε χρησιμοποιώντας τη βιβλιοθήκη mysql.connector και τη συνάρτηση create_engine από τη βιβλιοθήκη SQLAlchemy. Εξηγήθηκε λεπτομερώς η διαδικασία εκτέλεσης ερωτημάτων SQL και δημιουργίας πινάκων. Συνοψίζοντας, παρουσιάστηκε μια ολοκληρωμένη επισκόπηση της δημιουργίας συνδέσεων, της εκτέλεσης ερωτημάτων βάσης δεδομένων και του χειρισμού δεδομένων στην Python, αναδεικνύοντας την ευελιξία και τις δυνατότητες αυτών των εργαλείων για ένα ευρύ φάσμα εργασιών που σχετίζονται με βάσεις δεδομένων.

Επιπλέον, αξίζει να επισημανθεί ότι η καθιερωμένη φήμη της MySQL ως σημείο αναφοράς του κλάδου έχει δημιουργήσει μια πληθώρα πολύτιμων πόρων και ικανών προγραμματιστών, που εγγυώνται συνεχή πρόοδο και ισχυρή υποστήριξη των χρηστών. Η MySQL συνεχίζει να χρησιμεύει ως απαραίτητο εργαλείο για την αξιοποίηση του δυναμικού των δεδομένων με αποτελεσματικότητα, ταχύτητα και ασφάλεια. Η απόκτηση επάρκειας στη MySQL ξεκλειδώνει ένα βασίλειο ευκαιριών που βασίζονται στα δεδομένα, διευκολύνοντας τη δυναμική εξέλιξη της τεχνολογίας και του εμπορίου. [B23c].

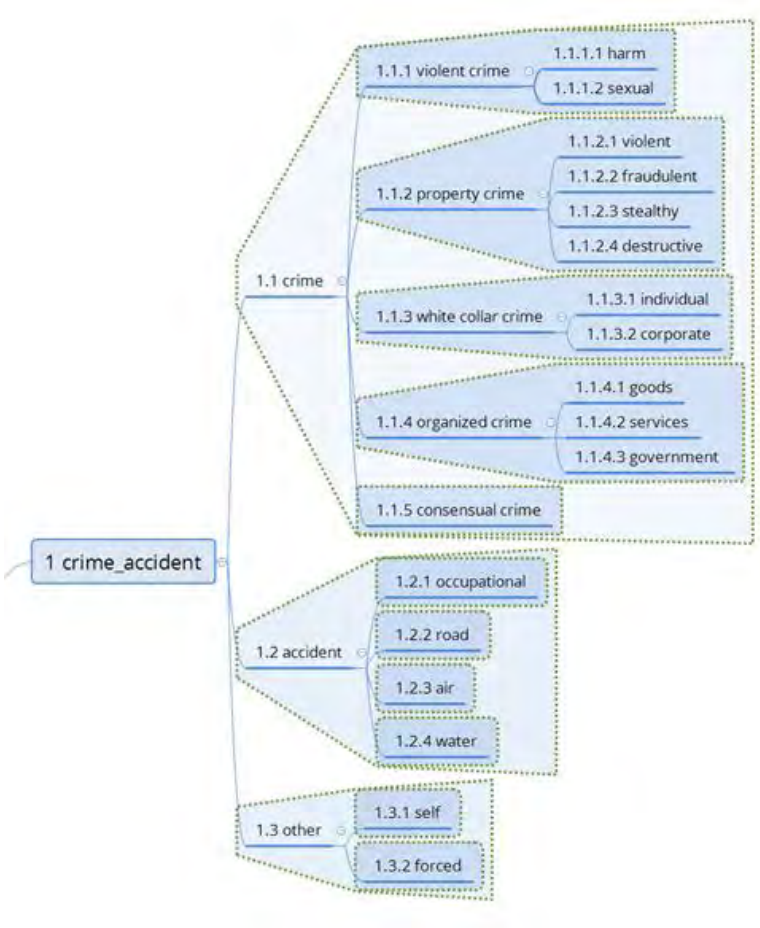
7 Δένδρο κατηγοριοποίησης ειδησεογραφικών γεγονότων



Εικόνα 32: Δένδρο κατηγοριοποίησης ειδησεογραφικών γεγονότων

Παραπάνω εμφανίζονται, μέσω της μορφής δενδροδιαγράμματος, οι κατηγορίες διαχωρισμού των ειδησεογραφικών γεγονότων στην προτεινόμενη βάση δεδομένων.

Κατηγορία Εγκλήματα – Ατυχήματα



Εικόνα 33: Δενδροδιάγραμμα κατηγορίας Εγκλημάτων - Ατυχημάτων

Η κατηγορία *(1) Εγκλήματα και Ατυχήματα*, διαχωρίστηκε αρχικά σε *(1.1) Εγκλήματα, (1.2) Ατυχήματα* και *(1.3) Άλλο είδος*. Ακολουθεί η θεωρητική ανάλυση της κάθε υποκατηγορίας.

Ως *(1.1) Έγκλημα* ορίζεται κάθε πράξη ή παράλειψη που συνιστά αδίκημα και τιμωρείται από το νόμο. Αυτή η υποκατηγορία χωρίζεται σε:

> *(1.1.1) Βίαια εγκλήματα* [Jus22]: Τα βίαια εγκλήματα ορίζονται ως εκείνα τα αδικήματα που περιλαμβάνουν βία ή απειλή βίας. Κατηγοριοποιούνται περαιτέρω σε:
>
> > *(1.1.1.1) Επιβλαβή εγκλήματα*: Εγκλήματα όπου ένα άτομο βλάπτει ή απειλεί να βλάψει άλλο άτομο (πχ. ανθρωποκτονία).

(1.1.1.2) Σεξουαλικά εγκλήματα: Τα προσβλητικά σεξουαλικά εγκλήματα κατηγοριοποιούνται ως βίαια εγκλήματα. Αυτά περιλαμβάνουν βιασμό και σεξουαλική επίθεση. Μπορεί να περιλαμβάνει βίαιη σωματική επαφή ή μπορεί να επιτευχθεί μέσω συναισθηματικής χειραγώγησης ή αδυναμίας του θύματος να συναινέσει. Σε οποιαδήποτε από αυτές τις συνθήκες, θεωρείται βίαιο έγκλημα λόγω των άκρως προσβλητικών παραβιάσεων που βιώνει το θύμα (πχ. βιασμός)

(1.1.2) Εγκλήματα ιδιοκτησίας: Έγκλημα ιδιοκτησίας είναι κάθε παράνομη δραστηριότητα που περιλαμβάνει την καταστροφή ή μεταβίβαση περιουσίας, ανεξάρτητα από το αν χρησιμοποιείται πράξη βίας [LRS19]. Κατηγοριοποιούνται περαιτέρω σε:

(1.1.2.1) Βίαιο έγκλημα ιδιοκτησίας: Περιλαμβάνει τον δράστη που χρησιμοποιεί ή απειλεί να χρησιμοποιήσει βία σε ένα θύμα κατά τη διάρκεια του συμβάντος (πχ. ληστεία).

(1.1.2.2) Έγκλημα δόλιας ιδιοκτησίας: Το έγκλημα δόλιας ιδιοκτησίας είναι η χρήση σκόπιμης εξαπάτησης για την εξασφάλιση παράνομων κερδών ή τη στέρηση των θυμάτων από τα νόμιμα δικαιώματά τους (πχ. υπεξαίρεση).

(1.1.2.3) Έγκλημα κλοπής ιδιοκτησίας: Ο νόμος ορίζει αυτό ως μη συναινετική, μη βίαιη και μη δόλια κλοπή περιουσίας. Για να θεωρηθεί το έγκλημα κρυφό, πρέπει να ολοκληρωθεί απουσία του θύματος (πχ. διάρρηξη).

(1.1.2.4) Έγκλημα καταστροφής ιδιοκτησίας: Αυτό το είδος ορίζεται ως η παράνομη καταστροφή περιουσίας (πχ. βανδαλισμός).

(1.1.3) Εγκλήματα «Whitecollar»: Το έγκλημα whitecollar είναι ένα μη βίαιο έγκλημα όπου το κύριο κίνητρο είναι συνήθως οικονομικό [CFI20]. Οι εγκληματίες του whitecollar κατέχουν συνήθως μια επαγγελματική θέση ισχύος ή/και κύρους, και μια θέση που αποζημιώνεται πολύ πάνω από το μέσο όρο. Κατηγοριοποιούνται περαιτέρω σε:

(1.1.3.1) Ατομικά: Τα ατομικά εγκλήματα είναι οικονομικά εγκλήματα που διαπράττονται από άτομο ή ομάδα ατόμων (πχ πλαστογραφία).

(1.1.3.2) Εταιρικά: Κάποιο έγκλημα του whitecollar που συμβαίνει σε εταιρικό επίπεδο (πχ. φοροδιαφυγή).

(1.1.4) Οργανωμένο έγκλημα: Το οργανωμένο έγκλημα είναι μια συνεχιζόμενη εγκληματική επιχείρηση που εργάζεται για να επωφεληθεί από παράνομες δραστηριότητες που συχνά έχουν μεγάλη δημόσια ζήτηση. Κατηγοριοποιούνται περαιτέρω σε:

(1.1.4.1) Παροχή παράνομων αγαθών (πχ. πορνογραφία).

(1.1.4.2) Παροχή παράνομων υπηρεσιών (πχ. πορνεία).

(1.1.4.3) Διείσδυση επιχειρήσεων και κυβέρνησης (πχ. τρομοκρατία).

(1.1.5) Έγκλημα με συγκατάθεση: Αναφέρεται σε συμπεριφορές στις οποίες οι άνθρωποι εμπλέκονται οικειοθελώς και πρόθυμα, παρόλο που αυτές οι συμπεριφορές παραβιάζουν το νόμο (πχ. οπλοκατοχή).

Ως *(1.2) Ατύχημα* ορίζεται ένα ατυχές περιστατικό που συμβαίνει απροσδόκητα και ακούσια, με αποτέλεσμα συνήθως ζημιά ή τραυματισμό [HSE22]. Αυτή η υποκατηγορία χωρίζεται σε:

(1.2.1) Ατυχήματα στο εργασιακό περιβάλλον (πχ. τραυματισμός).

(1.2.2) Ατυχήματα που συμβαίνουν στο δρόμο (πχ. τροχαίο).

(1.2.3) Ατυχήματα που συμβαίνουν στον εναέριο χώρο (πχ. αεροπορικό δυστύχημα).

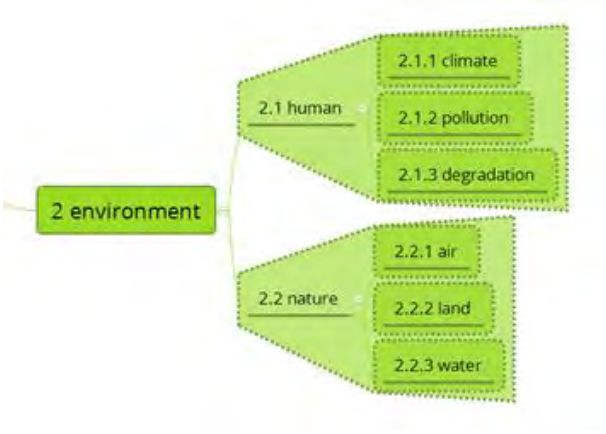
(1.2.4) Ατυχήματα που συμβαίνουν στο νερό (πχ. πετρελαιοκηλίδα).

Ως *(1.3) Άλλο είδος* ορίζονται ειδικές κατηγορίες εγκλημάτων-ατυχημάτων. Αυτή η υποκατηγορία χωρίζεται σε:

(1.3.1) Ειδικές κατηγορίες εγκλημάτων-ατυχημάτων που τα προκαλεί το ίδιο το άτομο στον εαυτό του (πχ. αυτοκτονία).

(1.3.2) Ειδικές κατηγορίες εγκλημάτων-ατυχημάτων που προκαλούνται από δεύτερο πρόσωπο (πχ. εξαφάνιση).

Κατηγορία Περιβάλλον



Εικόνα 34: Δενδροδιάγραμμα κατηγορίας Περιβάλλον

Η κατηγορία *(2) Περιβάλλον*, χωρίστηκε αρχικά σε:

(*2.1) Περιβαλλοντικές καταστροφές που οφείλονται σε ανθρώπινους παράγοντες* και κατηγοριοποιείται περαιτέρω σε:

(2.1.1.) Η ευθύνη του ανθρώπου όσον αφορά την *κλιματική αλλαγή* (πχ φαινόμενο του θερμοκηπίου).

(2.1.2) Η ευθύνη του ανθρώπου όσον αφορά την *μόλυνση του περιβάλλοντος* (πχ. καυσαέρια).

(2.1.3) Η ευθύνη του ανθρώπου όσον αφορά την *καταστροφή της χλωρίδας και της πανίδας* (πχ. πυρκαγιά).

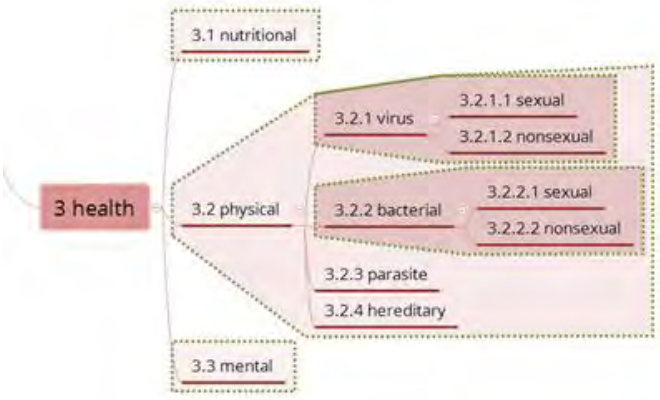
(2.2) Φυσικές καταστροφές και κατηγοριοποιείται περαιτέρω σε:

(2.2.1) Φυσικές καταστροφές όσον αφορά τον εναέριο χώρο (πχ. καταιγίδα).

(2.2.2) Φυσικές καταστροφές όσον αφορά τη γη. (πχ. σεισμός)

(2.2.3) Φυσικές καταστροφές όσον αφορά τον υδάτινο χώρο. (πχ. τσουνάμι)

Κατηγορία Υγεία



Εικόνα 35: Δενδροδιάγραμμα κατηγορίας Υγεία

Η κατηγορία *(3) Υγεία* αφορά γεγονότα που αφορούν την ανακάλυψη ενός νέου εμβολίου ή θεραπείας για μια ασθένεια και διαχωρίστηκε αρχικά σε *(3.1) Διατροφικές παθήσεις, (3.2) Σωματικές παθήσεις* και *(3.3) Ψυχολογικές – Διανοητικές παθήσεις*. Ακολουθεί η θεωρητική ανάλυση της κάθε υποκατηγορίας [NHS23].

Ως *(3.1) Διατροφικές παθήσεις* θεωρούνται οι ασθένειες που σχετίζονται με την διατροφή.

Ως *(3.2) Σωματικές παθήσεις* θεωρούνται οι ασθένειες που επηρεάζουν την φυσική κατάσταση του ανθρώπου. Κατηγοριοποιούνται περαιτέρω σε:

> *(3.2.1) Ιογενείς:* Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε ιό. Κατηγοριοποιούνται περαιτέρω σε:
>
> > *(3.2.1.1) Σεξουαλικό:* Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε σεξουαλικώς μεταδιδόμενο νόσημα/ιό (πχ. HIV).
> >
> > *(3.2.1.2) Μη σεξουαλικό:* Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και ευθύνεται σε μη σεξουαλικώς μεταδιδόμενο νόσημα/ιό (πχ. H1N1).
>
> *(3.2.2) Βακτηριάσεις*: Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε βακτήρια. Κατηγοριοποιούνται περαιτέρω σε:

(3.2.2.1) Σεξουαλικό: Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και ευθύνεται σε σεξουαλικώς μεταδιδόμενο νόσημα/βακτήριο (πχ. χλαμύδια).

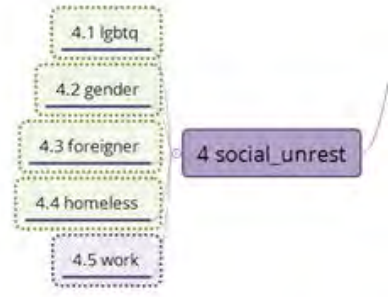
(3.2.2.2) Μη σεξουαλικό: Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε μη σεξουαλικώς μεταδιδόμενο νόσημα/βακτήριο (πχ. φυματίωση).

(3.2.3) Παρασιτογενείς: Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε παράσιτο (πχ. ψώρα).

(3.2.4) Κληρονομικές: Ό,τι επηρεάζει την φυσική κατάσταση του ανθρώπου και οφείλεται σε κληρονομικό παράγοντα (πχ. Άλτσχάιμερ).

Ως *(3.3) Ψυχολογικές – Διανοητικές παθήσεις* θεωρείται ό,τι επηρεάζει την πνευματική κατάσταση του ανθρώπου (πχ. Σχιζοφρένεια).

Κατηγορία Κοινωνικές Αναταραχές



Εικόνα 36: Δενδροδιάγραμμα κατηγορίας Κοινωνικές αναταραχές

Ως *(4) Κοινωνικές αναταραχές* ορίζονται οι πράξεις που οδηγούν σε έλλειψη τάξεως & ομαλής λειτουργίας της ζωής. Κατηγοριοποιούνται σε:

(*4.1) Αναταραχές της lgbtq κοινότητας* (πχ. pride).

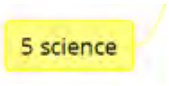
(4.2) Αναταραχές με βάση το φύλο (πχ. μισογυνισμός).

(4.3) Αναταραχές σχετικές με αλλοδαπούς, μετανάστες και ανθρώπους από άλλες χώρες (πχ. ρατσισμός).

(4.4) Αναταραχές που σχετίζονται με άστεγους (πχ. λιμός).

(4.5) Αναταραχές που σχετίζονται με τον χώρο εργασίας (πχ. απεργία).

Κατηγορία Επιστήμη



Εικόνα 37: Δενδροδιάγραμμα κατηγορίας Επιστήμη

Η κατηγορία *(5) Επιστήμη*, αφορά ανακαλύψεις και εφευρέσεις στον χώρο της επιστήμης (πχ. Ανακαλύφθηκε νέο υλικό που βοηθά τα διαβητικά τραύματα να επουλωθούν γρήγορα).

Κατηγορία Τέχνη

Εικόνα 38: Δενδροδιάγραμμα κατηγορίας Τέχνη

Η κατηγορία *(6) Τέχνη* αποτελείται από γεγονότα που αφορούν την τέχνη και τον πολιτισμό. Κατηγοριοποιείται σε:

(6.1) Μουσική: Γεγονότα που αφορούν την τέχνη της μουσικής (πχ. Φεστιβάλ Πόζαρ).

(6.2) Κινηματογράφος – Θέατρο: Γεγονότα που αφορούν την τέχνη στον τομέα του κινηματογράφου και του θεάτρου (πχ. Βραβεία Oscar).

(6.3) Μουσεία: Γεγονότα που αφορούν την τέχνη στον τομέα των μουσείων (πχ. Έκθεση ζωγραφικής).

(6.4) Ποίηση: Γεγονότα που αφορούν την τέχνη της ποίησης (πχ. Κρατικά Λογοτεχνικά Βραβεία).

Κατηγορία Ολυμπιακοί αγώνες

Εικόνα 39: Δενδροδιάγραμμα κατηγορίας Ολυμπιακοί αγώνες

Η κατηγορία *(7) Ολυμπιακοί αγώνες* αποτελείται από γεγονότα που αφορούν τα πιο γνωστά αθλήματα των ολυμπιακών αγώνων. Συγκεκριμένα, διακρίσεις και μετάλλια στα παρακάτω αθλήματα που κατηγοριοποιούνται σε:

(7.1) Θερινά: Τα πιο γνωστά θερινά αθλήματα των ολυμπιακών αγώνων. Χωρίζονται περαιτέρω σε:

(7.1.1) Υδάτινο περιβάλλον

(7.1.2) Κανό ή καγιάκ

(7.1.3) Ποδηλασία

(7.1.4) Βόλεϊ ή χαντμπολ

(7.1.5) Ιππασία

(7.1.6) Πάλη

(7.1.7) Γυμναστική

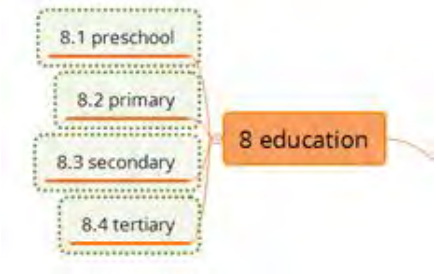
(7.2) Χειμερινά: Τα πιο γνωστά χειμερινά αθλήματα των ολυμπιακών αγώνων. Χωρίζονται περαιτέρω σε:

(7.2.1) Παγοδρόμιο

(7.2.2) Σκι

(7.2.3) Έλκηθρο

Κατηγορία Εκπαίδευση



Εικόνα 40: Δενδροδιάγραμμα κατηγορίας Εκπαίδευση

Η κατηγορία *(8) Εκπαίδευση* αποτελείται από γεγονότα που αφορούν την εκπαίδευση (πχ. Μεταρρύθμιση Παιδείας). Κατηγοριοποιείται σε:

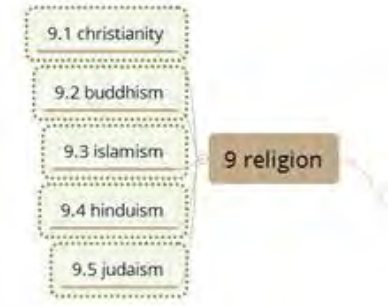
(8.1) Νηπιαγωγείο

(8.2) Πρωτοβάθμια εκπαίδευση

(8.3) Δευτεροβάθμια εκπαίδευση

(8.4) Τριτοβάθμια εκπαίδευση

Κατηγορία Θρησκεία



Εικόνα 41: Δενδροδιάγραμμα κατηγορίας Θρησκεία

Η κατηγορία *(9) Θρησκεία* αποτελείται από γεγονότα που αφορούν θρησκευτικά δρώμενα (πχ. Θάνατος θρησκευτικού ηγέτη). Κατηγοριοποιείται στις κυριότερες θρησκείες του κόσμου:

(9.1) Χριστιανισμός

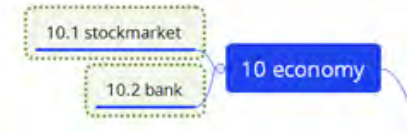
(9.2) Βουδισμός

(9.3) Ισλαμισμός

(9.4) Ινδουισμός

(9.5) Ιουδαϊσμός

Κατηγορία Οικονομία



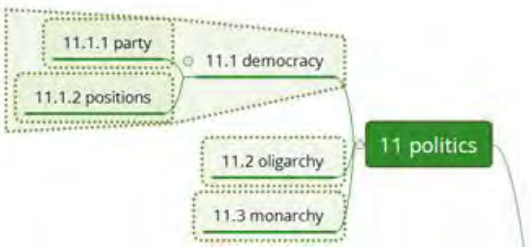
Εικόνα 42: Δενδροδιάγραμμα κατηγορίας Οικονομία

Η κατηγορία *(10) Οικονομία* αποτελείται από γεγονότα που αφορούν την οικονομία του κόσμου. Κατηγοριοποιείται σε:

(10.1) Χρηματιστήριο (πχ. πτώση χρηματιστηρίου).

(10.2) Τραπεζικός τομέας (πχ. κατάσχεση λόγω δανείου).

Κατηγορία Πολιτική



Εικόνα 43: Δενδροδιάγραμμα κατηγορίας Πολιτική

Η κατηγορία *(11) Πολιτική* αποτελείται από γεγονότα που αφορούν πολιτικά δρώμενα (πχ. εκλογές). Κατηγοριοποιείται στα κυριότερα πολιτεύματα:

(11.1) Δημοκρατία. Χωρίζεται περαιτέρω σε:

(11.1.1) Πολιτικά κόμματα

(11.1.2) Πολιτικές θέσεις/τάξεις

(11.2) Ολιγαρχία

(11.3) Μοναρχία

8 Αυτοματοποιημένη κατάταξη ειδησεογραφικών άρθρων βάσει περιεχομένου

8.1 Εισαγωγή

Σε έναν κόσμο κορεσμένο από πληροφορίες, η αποτελεσματική κατηγοριοποίηση των γεγονότων έχει καταστεί απαραίτητη. Είναι χρήσιμο ένα σύστημα που μπορεί γρήγορα να κατανοήσει τα άρθρα ειδήσεων, τις ενημερώσεις των μέσων κοινωνικής δικτύωσης και τις αναφορές αναλύοντας απλώς τις λέξεις-κλειδιά που περιέχουν. Αυτή είναι η ουσία της προσέγγισης κατηγοριοποίησης συμβάντων που βασίζεται σε λέξεις-κλειδιά.

Στο επίκεντρο αυτής της κατηγοριοποίησης βρίσκεται η δύναμη των λέξεων-κλειδιών. Αυτά τα συνοπτικά, γεμάτα πληροφορίες στοιχεία χρησιμεύουν ως τα θεμελιώδη δομικά στοιχεία της ταξινόμησης γεγονότων. Αναλύοντας τις συγκεκριμένες λέξεις και φράσεις που σχετίζονται με ένα γεγονός, μπορεί να γίνει αυτόματα η ταξινόμηση σε σχετικές κατηγορίες. Κατανοώντας, τις λέξεις-κλειδιά που σχετίζονται με ένα συμβάν, μπορούν να ξεκλειδωθούν ανεκτίμητες πληροφορίες σχετικά με τη φύση, το πλαίσιο και τη σημασία του.

Στη συνέχεια, απεικονίζονται οι λέξεις κλειδιά που χρησιμοποιήθηκαν για κάθε κατηγορία και υποκατηγορία γεγονότων.

8.2 Κατηγορία Εγκλημάτων-Ατυχημάτων (Crime-Accident)

(1) total_crime_accident_vocab

Λέξεις κλειδιά: 'έγκλημα', 'ατύχημα', 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'αυτοκτονία', 'δυστύχημα', 'τροχαίο', 'κλοπή', 'διάρρηξη', 'βασανισμός', 'ξυλοδαρμός', 'βιασμός', 'γυναικοκτονία', 'παρανομία', 'λαθρεμπόριο', 'πορνεία', 'εμπρησμός', 'τρομοκρατία', 'επίθεση', 'σεξουαλική παρενόχληση', 'διακίνηση', 'κακοποίηση', 'εξαφάνιση', 'απαγωγή', 'bullying', 'εκφοβισμός', 'εκβιασμός', 'ασέλγεια', 'καραμπόλα', 'πνιγμός', 'αποπλάνηση', 'απάτη', 'απαγχονισμός', 'πυροβολισμός', 'βανδαλισμός', 'γενοκτονία', 'διαφθορά', 'ληστεία', 'δυσφήμιση', 'φοροδιαφυγή', 'ναρκωτικά', 'amber alert', 'πτώμα', 'νεκρός', 'τραυματίας', 'δίκη', 'δικαστήριο', 'ποινή', 'φυλακή', 'καταγγελία', 'πορνογραφία' ,'δηλητηρίαση', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία', 'σωματεμπορία', 'πειρατεία', 'οπλοκατοχή', 'υπεξαίρεση', 'τραυματισμός', 'αεροπορικό', 'πετρελαιοκηλίδα', 'δουλεμπόριο', 'αρπαγή'

(1.1) total_crime_vocab

Λέξεις κλειδιά: 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'κλοπή', 'διάρρηξη', 'βασανισμός', 'ξυλοδαρμός', 'βιασμός', 'γυναικοκτονία', 'παρανομία', 'λαθρεμπόριο',

'πορνεία', 'εμπρησμός', 'τρομοκρατία', 'απαγχονισμός', 'επίθεση', 'σεξουαλική παρενόχληση', 'διακίνηση', 'κακοποίηση', 'απαγωγή', 'bullying', 'εκφοβισμός', 'εκβιασμός', 'ασέλγεια', 'αποπλάνηση', 'απάτη', 'πυροβολισμός', 'βανδαλισμός', 'γενοκτονία', 'διαφθορά', 'ληστεία', 'δυσφήμιση', 'φοροδιαφυγή', 'ναρκωτικά', 'πορνογραφία', 'δηλητηρίαση', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία', 'σωματεμπορία', 'πειρατεία', 'οπλοκατοχή', 'υπεξαίρεση', 'δουλεμπόριο', 'αρπαγή'

(1.1.1) violent_crime_vocab

Λέξεις κλειδιά: 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'βασανισμός', 'ξυλοδαρμός', 'βιασμός', 'γυναικοκτονία', 'απαγχονισμός', 'σεξουαλική παρενόχληση', 'κακοποίηση', 'ασέλγεια', 'δηλητηρίαση', 'αποπλάνηση', 'αρπαγή', 'απαγωγή'

(1.1.1.1) violent_harm_crime_vocab

Λέξεις κλειδιά: 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'βασανισμός', 'ξυλοδαρμός', 'γυναικοκτονία', 'απαγχονισμός', 'κακοποίηση', 'δηλητηρίαση', 'αρπαγή', 'απαγωγή'

(1.1.1.2) violent_sexual_crime_vocab

Λέξεις κλειδιά: 'βιασμός', 'σεξουαλική παρενόχληση', 'κακοποίηση', 'ασέλγεια', 'αποπλάνηση'

(1.1.2) property_crime_vocab

Λέξεις κλειδιά: 'κλοπή', 'διάρρηξη', 'εμπρησμός', 'βανδαλισμός', 'ληστεία', 'πειρατεία', 'υπεξαίρεση'

(1.1.2.1) property_violent_crime_vocab

Λέξεις κλειδιά: 'ληστεία', 'πειρατεία'

(1.1.2.2) property_fraudulent_crime_vocab

Λέξεις κλειδιά: 'υπεξαίρεση'

(1.1.2.3) property_stealthy_crime_vocab

Λέξεις κλειδιά: 'κλοπή', 'διάρρηξη'

(1.1.2.4) property_destructive_crime_vocab

Λέξεις κλειδιά: 'εμπρησμός', 'βανδαλισμός'

(1.1.3) whitecollar_crime_vocab

Λέξεις κλειδιά: 'απάτη', 'διαφθορά', 'φοροδιαφυγή', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία'

(1.1.3.1) whitecollar_individual_crime_vocab

Λέξεις κλειδιά: 'διαφθορά', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία'

(1.1.3.2) whitecollar_corporate_crime_vocab

Λέξεις κλειδιά: 'απάτη', 'διαφθορά', 'φοροδιαφυγή'

(1.1.4) organized_crime_vocab

Λέξεις κλειδιά: 'λαθρεμπόριο', 'πορνεία', 'τρομοκρατία', 'διακίνηση', 'πορνογραφία', 'σωματεμπορία', 'δουλεμπόριο'

(1.1.4.1) organized_goods_crime_vocab

Λέξεις κλειδιά: 'λαθρεμπόριο', 'διακίνηση', 'πορνογραφία'

(1.1.4.2) organized_services_crime_vocab

Λέξεις κλειδιά: 'πορνεία', 'σωματεμπορία', 'δουλεμπόριο'

(1.1.4.3) organized_government_crime_vocab

Λέξεις κλειδιά: 'τρομοκρατία'

(1.1.5) consensual_crime_vocab

Λέξεις κλειδιά: ‘πορνεία’, ‘ναρκωτικά’, ‘πορνογραφία’, ‘οπλοκατοχή’

(1.2) total_accident_vocab

Λέξεις κλειδιά: 'δυστύχημα', 'τροχαίο', 'καραμπόλα', 'πνιγμός', 'πυροβολισμός', 'τραυματισμός', 'αεροπορικό', 'πετρελαιοκηλίδα'

(1.2.1) accident_occupational_vocab

Λέξεις κλειδιά: 'τραυματισμός'

(1.2.2) accident_road_vocab

Λέξεις κλειδιά: 'δυστύχημα', 'τροχαίο', 'καραμπόλα'

(1.2.3) accident_air_vocab

Λέξεις κλειδιά: 'αεροπορικό'

(1.2.4) accident_water_vocab

Λέξεις κλειδιά: ‘πνιγμός’, ‘πετρελαιοκηλίδα’

(1.3) total_other_crime_accident_vocab

Λέξεις κλειδιά: ‘αυτοκτονία’, ‘απαγχονισμός’, ‘εξαφάνιση’, ‘amber alert’

(1.3.1) other_self_crime_accident_vocab

Λέξεις κλεδιά: 'αυτοκτονία', 'απαγχονισμός'

(1.3.2) other_forced_crime_accident_vocab

Λέξεις κλειδιά: 'εξαφάνιση', 'amber alert'

8.3 Κατηγορία Περιβάλλον (Environment)

(2) total_environment_vocab

Λέξεις κλειδιά: 'φωτιά', 'σεισμός', 'πλημμύρα', 'τσουνάμι', 'καταιγίδα', 'κεραυνός', 'θερμοκηπίου', 'υπερθέρμανση', 'λιώσιμο', 'ρύποι', 'ρύπανση', 'πυρκαγιά', 'απορρίμματα', 'καυσαέριο', 'πετρελαιοκηλίδα', 'κλιματική', 'ξηρασία'

Δεν θεωρούνται γεγονότα: ‘καιρός’, ‘βροχή’

(2.1) environment_human_vocab

Λέξεις κλειδιά: 'θερμοκηπίου', 'υπερθέρμανση', 'λιώσιμο', 'ρύποι', 'ρύπανση', 'πυρκαγιά', 'απορρίμματα', 'καυσαέριο', 'πετρελαιοκηλίδα', 'κλιματική', 'φωτιά'

(2.1.1) environment_human_climate_vocab

Λέξεις κλειδιά: 'θερμοκηπίου', 'υπερθερμανση', 'λιώσιμο', 'κλιματική'

(2.1.2) environment_human_pollution_vocab

Λέξεις κλειδιά: ‘ρύποι’, ‘ρύπανση’, ‘απορρίματα’, ‘καυσαέριο’, ‘πετρελαιοκηλίδα’

(2.1.3) environment_human_degradation_vocab

Λέξεις κλειδιά: 'πυρκαγιά', 'πετρελαιοκηλίδα', 'φωτιά'

(2.2) environment_nature_vocab

Λέξεις κλειδιά: ‘σεισμός’, ‘πλημμύρα’, ‘τσουνάμι’, ‘καταιγίδα’, ‘κεραυνός’, ‘ξηρασία’

(2.2.1) environment_nature_air_vocab

Λέξεις κλειδιά: ‘καταιγίδα’, ‘κεραυνός’

(2.2.2) environment_nature_land_vocab

Λέξεις κλειδιά: ‘σεισμός’, ‘ξηρασία’

(2.2.3) environment_nature_water_vocab

Λέξεις κλειδιά: 'πλημμύρα', 'τσουνάμι'

8.4 Κατηγορία Υγείας (Health)

(3) total_health_vocab

Λέξεις κλειδιά: 'covid-19', 'covid', 'SARS-CoV2', 'SARS-CoV1', 'κορονοϊός', 'γρίπη', 'κρούσματα', 'εμβόλιο', 'ιός', 'συνταγογράφηση', 'φάρμακα', 'μάσκα', 'HIV', 'νόσος', 'ιατρός', 'ασθένεια', 'καρκίνος', 'cancer', 'Αλτσχάιμερ', 'Alzheimer', 'Πάρκινσον', 'Parkinson', 'λοίμωξη', 'διάγνωση', 'H1N1', 'ΕΟΦ', 'σχιζοφρένεια', 'διπολισμός', 'κατάθλιψη', 'άγχος', 'ΔΕΠΥ', 'μετατραυματικό', 'ψυχαναγκασμός', 'παχυσαρκία', 'ανορεξία', 'βουλιμία', 'αυτοάνοσο', 'υπέρταση', 'υπόταση', 'μυκητίαση', 'AIDS', 'εμπόλα', 'φυματίωση', 'ψώρα', 'διαβήτης', 'HPV', 'καρδιαναγγειακή', 'ηπατίτιδα', 'οστεοπόρωση', 'ανεύρυσμα', 'μονοπυρήνωση', 'γονόρροια', 'χλαμύδια', 'σύφιλη', 'Huntington'

(3.1) total_health_nutritional_vocab

Λέξεις κλειδιά: 'παχυσαρκία', 'ανορεξία', 'βουλιμία', 'διαβήτης'

(3.2) total_health_physical_vocab

Λέξεις κλειδιά: 'covid-19', 'covid', 'SARS-CoV2', 'SARS-CoV1', 'κορονοϊός', 'γρίπη', 'εμβόλιο', 'ιός', 'HIV', 'νόσος', 'ασθένεια', 'καρκίνος', 'cancer', 'Αλτσχάιμερ', 'Alzheimer',
'Πάρκινσον', 'Parkinson', 'λοίμωξη', 'H1N1', 'αυτοάνοσο', 'υπέρταση', 'υπόταση', 'μυκητίαση', 'AIDS', 'εμπόλα', 'φυματίωση', 'ψώρα', 'HPV', 'καρδιοαγγειακή', 'ηπατίτιδα', 'οστεοπόρωση', 'ανεύρυσμα', 'μονοπυρήνωση', 'γονόρροια', 'χλαμύδια',ν 'σύφιλη', 'Huntington'

(3.2.1) total_health_physical_virus_vocab

Λέξεις κλειδιά: 'covid-19', 'covid', 'SARS-CoV2', 'SARS-CoV1', 'κορονοϊός', 'γρίπη', 'ιός', 'HIV', 'H1N1', 'AIDS', 'εμπόλα', 'HPV', 'ηπατίτιδα', 'μονοπυρήνωση'

(3.2.1.1) total_health_physical_virus_sexual_vocab

Λέξεις κλειδιά: 'HIV', 'AIDS', 'HPV', 'ηπατίτιδα', 'μονοπυρήνωση'

(3.2.1.2) total_health_physical_virus_nonsexual_vocab

Λέξεις κλειδιά: 'covid-19', 'covid', 'SARS-CoV2', 'SARS-CoV1', 'κορονοϊός', 'γρίπη', 'H1N1', 'εμπόλα'

(3.2.2) total_health_physical_bacterial_vocab

Λέξεις κλειδιά: 'φυματίωση', 'γονόρροια', 'χλαμύδια', 'σύφιλη'

(3.2.2.1) total_health_physical_bacterial_sexual_vocab

Λέξεις κλειδιά: 'γονόρροια', 'χλαμύδια', 'σύφιλη'

(3.2.2.2) total_health_physical_bacterial_nonsexual_vocab

Λέξεις κλειδιά: 'φυματίωση'

(3.2.3) total_health_physical_parasite_vocab

Λέξεις κλειδιά: 'ψώρα'

(3.2.4) total_health_physical_hereditary_vocab

Λέξεις κλειδιά: 'καρκίνος', 'cancer', 'Αλτσχάιμερ', 'Alzheimer', 'Πάρκινσον', 'Parkinson', 'αυτοάνοσο', 'καρδιοαγγειακή', 'οστεοπόρωση', 'ανεύρυσμα', 'Huntington'

(3.3) total_health_mental_vocab

Λέξεις κλειδιά: 'σχιζοφρένεια', 'διπολισμός', 'κατάθλιψη', 'άγχος', 'ΔΕΠΥ', 'μετατραυματικό', 'ψυχαναγκασμός'

8.5 Κατηγορία Κοινωνικών Αναταραχών (Social Unrest)

(4) total_social_unrest_vocab

Λέξεις κλειδιά: 'διαδήλωση', 'πορεία', 'κινητοποίηση', 'συλλαλητήριο', 'LGBTQ', 'ισότητα', 'ΛΟΑΤΚΙ', 'ρατσισμός', 'μισογυνισμός', 'μισανδρία', 'πατριαρχία', 'μετανάστες', 'αλλοδαποί', 'πρόσφυγες', 'άστεγοι', 'λιμός', 'τριτοκοσμική', 'ΟΗΕ', 'pride', 'μητριαρχία', 'απεργία', 'ομοφοβία'

(4.1) total_social_unrest_lgbtq_vocab

Λέξεις κλειδιά: 'LGBTQ', 'ΛΟΑΤΚΙ', 'ομοφοβία', 'pride'

(4.2) total_social_unrest_gender_vocab

Λέξεις κλειδιά: 'ισότητα', 'μισογυνισμός', 'μισανδρία', 'πατριαρχία', 'μητριαρχία'

(4.3) total_social_unrest_foreigner_vocab

Λέξεις κλειδιά: 'ρατσισμός', 'μετανάστες', 'αλλοδαποί', 'πρόσφυγες', 'τριτοκοσμική', 'ΟΗΕ'

(4.4) total_social_unrest_homeless_vocab

Λέξεις κλειδιά: 'άστεγοι', 'λιμός'

(4.5) total_social_unrest_work_vocab

Λέξεις κλειδιά: 'απεργία'

8.6 Κατηγορία Επιστήμη (Science)

(5) total_science_vocab

Λέξεις κλειδιά: 'τεχνολογία', 'καινοτομία', 'ρομπότ', 'επιστήμη', 'ανακάλυψη', 'εφεύρεση', 'γενετική', 'μεθοδολογία', 'νόμπελ', 'science'

8.7 Κατηγορία Τέχνη (Arts)

(6) total_arts_vocab

Λέξεις κλειδιά: 'τέχνη', 'πολιτισμός', 'μουσείο', 'gallery', 'γκαλερί', 'πινακοθήκη', 'μουσική', 'ζωγραφική', 'συναυλία', 'εκδήλωση', 'έκθεση', 'θέατρο', 'κινηματογράφος', 'ταινία', 'Όσκαρ', 'φεστιβάλ', 'ποίηση', 'ποιητής', 'γλυπτά', 'άγαλμα'

(6.1) total_arts_music_vocab

Λέξεις κλειδιά: 'μουσική', 'συναυλία', 'φεστιβάλ'

(6.2) total_arts_movie_theatre_vocab

Λέξεις κλειδιά: 'θέατρο', 'κινηματογράφος', 'ταινία', 'Όσκαρ'

(6.3) total_arts_museum_vocab

Λέξεις κλειδιά**:** 'μουσείο', 'gallery', 'γκαλερί', 'πινακοθήκη', 'ζωγραφική', 'έκθεση', 'γλυπτά', 'άγαλμα'

(6.4) total_arts_poetry_vocab

Λέξεις κλειδιά: 'ποίηση', 'ποιητής'

8.8 Κατηγορία Ολυμπιακών Αγώνων (Olympic Games)

(7) total_olympic_games_vocab

Λέξεις κλειδιά: 'κολύμβηση', 'υδατοσφαίριση', 'κανόε', 'καγιάκ', 'ποδηλασία', 'ενόργανη', 'ρυθμική', 'τραμπολίνο', 'πετοσφαίριση', 'βόλεϊ', 'πάλη', 'τοξοβολία', 'στίβος', 'αντιπτέριση', 'τένις', 'καλαθοσφαίριση', 'πυγμαχία', 'ξιφασκία', 'γκολφ', 'χειροσφαίριση', 'χάντμπολ', 'πένταθλο', 'κωπηλασία', 'ιστιοπλοΐα', 'σκοποβολή', 'τρίαθλο', 'ιππασία', 'κρίκοι', 'παγοδρομία', 'πατινάζ', 'χόκεϊ', 'κέρλινγκ', 'σκι', 'έλκηθρο'

Δεν θεωρούνται γεγονότα: ‘μετάλλιο’, ‘νίκη’, ΄πρώτη’, ‘δεύτερη’, ‘τρίτη’, ‘θέση’, ‘διάκριση’

(7.1) total_olympic_games_solstice_vocab

Λέξεις κλειδιά: 'κολύμβηση', 'υδατοσφαίριση', 'κανόε', 'καγιάκ', 'ποδηλασία', 'ενόργανη', 'ρυθμική', 'τραμπολίνο', 'πετοσφαίριση', 'βόλεϊ', 'πάλη', 'τοξοβολία', 'στίβος', 'αντιπτέριση', 'τένις', 'καλαθοσφαίριση', 'πυγμαχία', 'ξιφασκία', 'γκολφ', 'χειροσφαίριση', 'χαντμπολ', 'πένταθλο', 'κοπηλασία', 'ιστιοπλοΐα', 'σκοποβολή', 'τρίαθλο', 'ιππασία'

(7.1.1) total_olympic_games_solstice_wet_track_vocab

Λέξεις κλειδιά: 'κολύμβηση', 'υδατοσφαίριση'

(7.1.2) total_olympic_games_solstice_canoe_vocab

Λέξεις κλειδιά: 'κανόε', 'καγιάκ'

(7.1.3) total_olympic_games_solstice_cycling_vocab

Λέξεις κλειδιά: 'ποδηλασία'

(7.1.4) total_olympic_games_solstice_handball_vocab

Λέξεις κλειδιά: 'πετοσφαίριση', 'βόλεϊ', 'χειροσφαίριση', 'χαντμπολ'

(7.1.5) total_olympic_games_solstice_riding_vocab

Λέξεις κλειδιά: 'ιππασία'

(7.1.6) total_olympic_games_solstice_wrestling_vocab

Λέξεις κλειδιά: 'πάλη'

(7.1.7) total_olympic_games_solstice_gymnastics_vocab

Λέξεις κλειδιά: 'ενόργανη', 'ρυθμική', 'τραμπολίνο', 'κρίκοι'

(7.2) total_olympic_games_winter_vocab

Λέξεις κλειδιά: 'παγοδρομία', 'πατινάζ', 'χόκεϊ', 'κέρλινγκ', 'σκι', 'έλκηθρο'

(7.2.1) total_olympic_games_winter_ice_skating_vocab

Λέξεις κλειδιά: 'παγοδρομία', 'πατινάζ', 'χόκεϊ', 'κέρλινγκ'

(7.2.2) total_olympic_games_winter_ski_vocab

Λέξεις κλειδιά: 'σκι'

(7.2.3) total_olympic_games_winter_sleigh_vocab

Λέξεις κλειδιά: 'έλκηθρο'

8.9 Κατηγορία Εκπαίδευση (Education)

(8) total_education_vocab

Λέξεις κλειδιά: 'εκπαίδευση', 'γυμνάσιο', 'λύκειο', 'πανεπιστήμιο', 'νηπιαγωγείο', 'ΙΕΚ', 'ΑΕΙ', 'παιδεία', 'συγγράμματα', 'φοιτητής', 'μαθητής', 'δάσκαλος', 'πανελλήνιες', 'μηχανογραφικό', 'φροντιστήριο', 'πρωτοβάθμια', 'δευτεροβάθμια', 'τριτοβάθμια'

(8.1) total_education_preschool_vocab

Λέξεις κλειδιά: 'νηπιαγωγείο'

(8.2) total_education_primary_vocab

Λέξεις κλειδιά: 'δάσκαλος', 'πρωτοβάθμια'

(8.3) total_education_secondary_vocab

Λέξεις κλειδιά: 'γυμνάσιο', 'λύκειο', 'πανελλήνιες', 'μηχανογραφικό', 'φροντιστήριο', 'δευτεροβάθμια'

(8.4) total_education_tertiary_vocab

Λέξεις κλειδιά: 'πανεπιστήμιο', 'ΙΕΚ', 'ΑΕΙ', 'συγγράμματα', 'φοιτητής', 'τριτοβάθμια'

8.10 Κατηγορία Θρησκεία (Religion)

(9) total_religion_vocab

Λέξεις κλειδιά: 'θρησκεία', 'ανεξιθρησκεία', 'χριστιανισμός', 'μουσουλμανισμός', 'ινδουισμός', 'βουδισμός', 'αρχιερέας', 'επίσκοπος', 'μητροπολίτης', 'ιερέας', 'Ιερουσαλήμ', 'Βούδας', 'Κοράνι', 'εκκλησία', 'τζαμί', 'προσκύνημα', 'κηδεία', 'ορθοδοξία', 'καθολισμός', 'δόγμα', 'ισλαμισμός', 'ραμαντάν', 'χαλίφης', 'ιμάμης', 'γιόγκα', 'ιουδαϊσμός', 'ραβίνος', 'ιεχωβά'

Δεν θεωρούνται γεγονότα: 'γιορτή', 'εορτή', 'Πάσχα', 'Χριστούγεννα'

(9.1) total_religion_christianity_vocab

Λέξεις κλειδιά: 'χριστιανισμός', 'αρχιερέας', 'επίσκοπος', 'μητροπολίτης', 'ιερέας', 'εκκλησία', 'ορθοδοξία', 'καθολισμός'

(9.2) total_religion_buddhism_vocab

Λέξεις κλειδιά: 'βουδισμός', 'Βούδας'

(9.3) total_religion_islamism_vocab

Λέξεις κλειδιά: 'μουσουλμανισμός', 'Κοράνι', 'τζαμί', 'ισλαμισμός', 'ραμαντάν', 'χαλίφης', 'ιμάμης'

(9.4) total_religion_hinduism_vocab

Λέξεις κλειδιά: 'ινδουισμός', 'δέδες', 'γιόγκα'

(9.5) total_religion_judaism_vocab

Λέξεις κλειδιά: 'ιουδαϊσμός', 'ραβίνος', 'ιεχωβά'

8.11 Κατηγορία Οικονομία (Economy)

(10) total_economy_vocab

Λέξεις κλειδιά: 'χρηματιστήριο', 'μετοχές', 'τράπεζα', 'χρήματα', 'πιστωτική', 'χρεωστική', 'δάνειο', 'χρέος', 'υποθήκη', 'πτώχευση', 'επιδότηση', 'κρυπτονόμισμα', 'επένδυση', 'επιχείρηση'

Δεν θεωρούνται γεγονός: 'κέρδισε'

(10.1) total_economy_stockmarket_vocab

Λέξεις κλειδιά: 'χρηματιστήριο', 'μετοχές', 'κρυπτονόμισμα', 'επένδυση'

(10.2) total_economy_bank_vocab

Λέξεις κλειδιά: 'τράπεζα', 'πιστωτική', 'χρεωστική', 'δάνειο', 'χρέος', 'υποθήκη', 'επιχείρηση'

8.12 Κατηγορία Πολιτική (Politics)

(11) total_politics_vocab

Λέξεις κλειδιά: 'κόμμα', 'σύριζα', 'ΝΔ', 'ΚΚΕ', 'δημοκρατία', 'ΠΑΣΟΚ', 'ΚΙΝΑΛ', 'ΕΚ', 'ΕΛ', 'ΑΝΤΑΡΣΥΑ', 'ΟΚΔΕ', 'ΕΣΥ', 'ΕΕΚ', 'ΛΑΕ', 'ΜΕΡΑ25', 'ΔΗΞΑΝ', 'ΕΠΑΜ', 'ΑΚΚΕΛ', 'κοινοβούλιο', 'βουλή', 'ΔΗΜΑΡ', 'εκλογές', 'πρωθυπουργός', 'υπουργός', 'κυβέρνηση', 'βουλευτής', 'νομοθεσία', 'υφυπουργός', 'ολιγαρχία', 'πολίτευμα', 'φασισμός', 'ναζί', 'μοναρχία', 'δικτατορία', 'κομμουνισμός', 'κράτος', 'ΝΙΚΗ', 'Σπαρτιάτες'

Δεν θεωρούνται γεγονότα: 'ομιλία', 'επίσκεψη', 'συνάντηση'

(11.1) total_politics_democracy_vocab

Λέξεις κλειδιά**:** 'κόμμα', 'δημοκρατία', 'κοινοβούλιο', 'βουλή', 'εκλογές', 'πρωθυπουργός', 'υπουργός', 'βουλευτής', 'υφυπουργός', 'κομμουνισμός', 'ΝΔ', 'ΚΚΕ', 'ΧΑ', 'ΕΚ', 'ΕΛ', 'ΑΝΤΑΡΣΥΑ', 'ΟΚΔΕ', 'ΕΣΥ', 'ΕΕΚ', 'ΛΑΕ', 'ΜΕΡΑ25', 'ΔΗΞΑΝ', 'ΕΠΑΜ', 'ΑΚΚΕΛ', 'ΠΑΣΟΚ', 'ΚΙΝΑΛ', 'ΟΠ', 'ΑΝΕΛ', 'σύριζα', 'ΝΙΚΗ', 'Σπαρτιάτες'

(11.1.1) total_politics_democracy_party_vocab

Λέξεις κλειδιά: 'ΝΔ', 'ΚΚΕ', 'ΧΑ', 'ΕΚ', 'ΕΛ', 'ΑΝΤΑΡΣΥΑ', 'ΟΚΔΕ', 'ΕΣΥ', 'ΕΕΚ', 'ΛΑΕ', 'ΜΕΡΑ25', 'ΔΗΞΑΝ', 'ΕΠΑΜ', 'ΑΚΚΕΛ', 'ΠΑΣΟΚ', 'ΚΙΝΑΛ', 'ΟΠ', 'ΑΝΕΛ', 'σύριζα', 'ΝΙΚΗ', 'Σπαρτιάτες'

(11.1.2) total_politics_democracy_positions_vocab

Λέξεις κλειδιά: 'πρωθυπουργός', 'υπουργός', 'βουλευτής', 'υφυπουργός'

(11.2) total_politics_oligarchy_vocab

Λέξεις κλειδιά: 'ολιγαρχία', 'δικτατορία', 'φασισμός', 'ναζί'

(11.3) total_politics_monarchy_vocab

Λέξεις κλειδιά: 'μοναρχία'

Στη συνέχεια αναλύονται οι συντομογραφίες των κομμάτων.

Πίνακας 1: Πίνακας συντομογραφιών των πιο γνωστών ελληνικών πολιτικών κομμάτων

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ ΚΟΜΜΑΤΩΝ	
ΝΔ	Νέα Δημοκρατία
ΚΚΕ	Κομμουνιστικό Κόμμα Ελλάδας
ΕΚ	Ένωση Κεντρώων
ΕΛ	Ελληνική Λύση
ΑΝΤΑΡΣΥΑ	Αντικαπιταλιστική Αριστερή Συνεργασία για την Ανατροπή
ΟΚΔΕ	Οργάνωση Κομμουνιστών Διεθνιστών Ελλάδας
ΕΣΥ	Ελλήνων Συνέλευσις
ΕΕΚ	Εργατικό Επαναστατικό Κόμμα
ΛΑΕ	Λαϊκή Ενότητα
ΜΕΡΑ25	Μέτωπο Ευρωπαϊκής Ρεαλιστικής Ανυπακοής
ΔΗΞΑΝ	Δημιουργία, Ξανά!
ΕΠΑΜ	Ενιαίο Παλλαϊκό Μέτωπο
ΑΚΚΕΛ	Αγροτικό Κτηνοτροφικό Κόμμα Ελλάδας
ΠΑΣΟΚ	Πανελλήνιο Σοσιαλιστικό Κίνημα
ΟΠ	Οικολόγοι Πράσινοι
ΑΝΕΛ	Ανεξάρτητοι Έλληνες
ΣΥΡΙΖΑ	Συνασπισμός Ριζοσπαστικής Αριστεράς
ΚΙΝΑΛ	Κίνημα Αλλαγής
ΧΑ	Χρυσή Αυγή
ΝΙΚΗ	-
Σπαρτιάτες	-
Πλεύση ελευθερίας	-

8.13 Συνάρτηση categorization

Η χρήση των λέξεων-κλειδιών σε κάθε κατηγορία χρησιμοποιούνται στη συνάρτηση **categorization** η οποία καλείται σε όλες τις κατηγορίες και υποκατηγορίες προκειμένου να καταταχθεί το κάθε γεγονός στην κατηγορία που ανήκει κανονικά σύμφωνα με το περιεχόμενό του. Έχει σχεδιαστεί για να επεξεργάζεται και να κατηγοριοποιεί τα άρθρα ειδήσεων με βάση τους τίτλους και τις περιγραφές τους. Ενσωματώνει διάφορα βήματα για να επιτευχθεί αυτό, συμπεριλαμβανομένης της προεπεξεργασίας, του φιλτραρίσματος ομοιότητας, της αφαίρεσης διπλότυπων και της

κατηγοριοποίησης. Παρουσιάζεται πρώτα η συνάρτηση όπως είναι στον κώδικα, και μετέπειτα αναλύεται.

```
# Κατηγοριοποίηση των ειδήσεων
def categorization(total_vocab, title_totalvocab_stemmed, title_total_indexes, descr_totalvocab_stemmed,
                   descr_total_indexes):
    # δημιουργία dataframe total_vocab_frame όπου κάθε γραμμή περιέχει την λέξη που έχει υποστεί stemming και την λέξη χωρίς stemming
    total_vocab_stemmed = []  # λίστα με όλα τα tokens που έχουν υποστεί stemming
    total_vocab_tokenized = []  # λίστα με όλα τα tokens
    total_indexes = []
    k = 0

    for i in total_vocab:

        allwords_stemmed = tokenize_and_stem(i)  # για κάθε λέξη της total_vocab, tokenize και stemming
        total_vocab_stemmed.extend(allwords_stemmed)  # επέκταση της total_vocab_stemmed λίστας

        allwords_tokenized = tokenize_only(i)  # για κάθε λέξη της total_vocab, tokenize
        total_vocab_tokenized.extend(allwords_tokenized)  # επέκταση της total_vocab_tokenized λίστας

        for j in allwords_stemmed:
            total_indexes.append(k)

        k = k + 1
    # δημιουργία του dataframe total_vocab_frame
    total_vocab_frame = pd.DataFrame(
        {'stemmed': total_vocab_stemmed, 'words': total_vocab_tokenized},
        index=total_indexes)
    #display(total_vocab_frame)

    # Στο total_df μπαίνουν οι ειδήσεις συμφωνα με τα title, description
    total_df = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])
    total_dft = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])
    total_dfd = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])
```

```
# total_dft
count = 0
for i in title_totalvocab_stemmed:
    for j in total_vocab_stemmed:
        if i == j:
            # print(str(title_total_indexes[count]) + "  " + i)
            total_df.loc[df.index[title_total_indexes[count]]] = df.iloc[title_total_indexes[count]]
            try:
                total_dft = pd.concat(
                    [total_df, total_dft])  # συνένωση με το dataframe total_dft της προηγούμενης επανάληψης
            except:
                total_dft = total_df
    count = count + 1
total_dft = total_dft.drop_duplicates(subset="title")

# total_dfd
count = 0
count_index = 0
for i in descr_totalvocab_stemmed:
    for j in total_vocab_stemmed:
        if i == j:
            #print(str(descr_total_indexes[count]) + "  " + i)
            if descr_total_indexes[count] == descr_total_indexes[count - 1]:
                count_index = count_index + 1
                if count_index > 2:
                    total_df.loc[df.index[descr_total_indexes[count]]] = df.iloc[descr_total_indexes[count]]
                    try:
                        total_dfd = pd.concat(
                            [total_df, total_dfd])  # συνένωση με το dataframe total_dfd της προηγούμενης επανάληψης
                    except:
                        total_dfd = total_df
    count = count + 1
total_dfd = total_dfd.drop_duplicates(subset="description")

# συνένωση των dataframes total_dft, total_dfd
frames = [total_dft, total_dfd]
total_df = pd.concat(frames)
total_df = total_df.drop_duplicates(subset="title")
```

```
# total_df.reset_index(drop=True, inplace=True)
# pd.set_option('display.max_columns', None)
# pd.set_option('display.max_rows', None)
# display(total_df)

# εύρεση όμοιων ειδήσεων και αφαίρεσή τους
total_title_list = []
for title in total_df['title']:
    total_title_list.append(title)
# display(total_title_list)

similarity_filter(total_title_list, 0.725)
remove_duplicates(total_title_list, 0.9)

# Στο total_df3 μπαίνουν οι τελικές ειδήσεις, αφαίρεση όμοιων ειδήσεων με βάση τον τίτλο
total_df2 = pd.DataFrame(columns=['id',  'title', 'category', 'description', 'link', 'content', 'pubDate'])
total_df3 = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])

for i, j in enumerate(total_df['title']):
    for l, k in enumerate(total_title_list):
        if j == k:
            total_df2 = total_df.loc[total_df['title'] == k]  # επιλογή των όμοιων ειδήσεων
            total_df3 = pd.concat([total_df3, total_df2])

# total_df3.reset_index(drop=True, inplace=True)
# pd.set_option('display.max_columns', None)
# pd.set_option('display.max_rows', None)
# display(total_df3)

total_description_list = []
for description in total_df['description']:
    total_description_list.append(description)
# display(total_description_list)
# os.environ['SPACY_WARNING_IGNORE'] = 'W008'

similarity_filter(total_description_list, 0.82)
```

```
# Στο total_df5 μπαίνουν οι τελικές ειδήσεις, αφαίρεση όμοιων ειδήσεων με βάση το περιεχόμενο
total_df4 = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])
total_df5 = pd.DataFrame(columns=['id', 'title', 'category', 'description', 'link', 'content', 'pubDate'])

for i, j in enumerate(total_df3['description']):
    for l, k in enumerate(total_description_list):
        if j == k:
            total_df4 = total_df3.loc[total_df3['description'] == k]  # επιλογή των όμοιων ειδήσεων
            total_df5 = pd.concat([total_df5, total_df4])

# total_df5.reset_index(drop=True, inplace=True)
# pd.set_option('display.max_columns', None)
# pd.set_option('display.max_rows', None)
# display(total_df5)

return total_df5
```

Εικόνα 44: Συνάρτηση categorization

Η συνάρτηση κατηγοριοποίησης εκτελεί τα ακόλουθα βήματα για την επεξεργασία και την κατηγοριοποίηση άρθρων ειδήσεων:

1. Αρχικοποίηση δεδομένων: Γίνεται αρχικοποίηση τριών κενών DataFrames: total_df, total_dft και total_dfd.
2. Επεξεργασία τίτλων ειδήσεων: Επανάληψη κάθε λέξης με βάση το title_totalvocab_stemmed και αντιστοίχισή της με το λεξιλόγιο. Όταν βρεθεί μια αντιστοίχιση, προστίθονται τα αντίστοιχα δεδομένα ειδήσεων στο total_df και συνδέεται με το total_dft.
3. Αφαίρεση διπλότυπων τίτλων: Κατάργηση των διπλότυπων εγγραφών από το total_dft DataFrame με βάση τη στήλη "title".

4. Επεξεργασία Περιγραφών Ειδήσεων: Επανάληψη κάθε βασικής λέξης στο descr_totalvocab_stemmed και αντιστοίχισή της με το λεξιλόγιο. Όταν βρεθεί μια αντιστοίχιση, προστίθονται τα αντίστοιχα δεδομένα ειδήσεων στο total_df και συνδέονται με το total_dfd. Γίνεται διαχείριση πιθανών επαναλαμβανόμενων καταχωρήσεων για να εξασφαλιστεί η σωστή συνένωση.
5. Κατάργηση διπλότυπων περιγραφών: Κατάργηση των διπλότυπων εγγραφών από το total_dfd DataFrame με βάση τη στήλη "description".
6. Συγχώνευση DataFrames: Συνδυάζονται τα total_dft και total_dfd DataFrame σε ένα ενιαίο DataFrame με το όνομα total_df.
7. Δημιουργία λιστών για φιλτράρισμα ομοιότητας: Δημιουργούνται δύο λίστες: total_title_list για αποθήκευση τίτλων ειδήσεων και total_description_list για αποθήκευση περιγραφών ειδήσεων.
8. Φιλτράρισμα για την εύρεση ομοιότητας και αφαίρεση διπλότυπων: Γίνεται εφαρμογή συναρτήσεων που αναλύονται από την Ρ. Μάντζαρη στο κεφάλαιο 7 της εργασίας της [M23].
9. Δημιουργία τελικών DataFrames- Μέρος 1: Αρχικοποίηση των total_df2 και total_df3 DataFrames. Επανάληψη του total_df και του total_title_list για την εύρεση αντίστοιχων τίτλων ειδήσεων και την προσθήκη τους στο total_df2. Έπειτα γίνεται σύνδεση των ειδήσεων που προκύπτουν με το total_df3.
10. Δημιουργία λιστών για φιλτράρισμα ομοιότητας (περιγραφή): Δημιουργία μιας λίστας με το όνομα total_description_list για την αποθήκευση των περιγραφών των ειδήσεων.
11. Φιλτράρισμα ομοιότητας για περιγραφές: Γίνεται εφαρμογή συνάρτησης που περιγράφεται στο κεφάλαιο 7 της εργασίας [M23].
12. Δημιουργία τελικών DataFrames- Μέρος 2: Αρχικοποίηση των total_df4 και total_df5 DataFrames. Επανάληψη του total_df3 και του total_description_list για την εύρεση αντίστοιχων περιγραφών των ειδήσεων και να την προσθήκη τους στο total_df4. Έπειτα γίνεται σύνδεση των ειδήσεων που προκύπτουν με το total_df5.
13. Επιστρέφοντας το τελικό αποτέλεσμα: Επιστρέφεται το total_df5 DataFrame, το οποίο περιέχει τα τελικά επεξεργασμένα και κατηγοριοποιημένα δεδομένα ειδήσεων.

Συνολικά, η λειτουργία κατηγοριοποίησης εκτελεί μια σειρά βημάτων επεξεργασίας δεδομένων, συμπεριλαμβανομένου του φιλτραρίσματος παρόμοιων και διπλότυπων ειδήσεων, σύμφωνα με τις Ενότητες 7.1.2 και 7.1.3, αντιστοίχως, στην εργασία [M23], με αποτέλεσμα ένα κατηγοριοποιημένο DataFrame που διατηρεί τα επεξεργασμένα δεδομένα ειδήσεων. Η λειτουργία δείχνει τη χρήση διαφόρων βοηθητικών λειτουργιών για την επίτευξη του σκοπού της.

Η κλήση της συνάρτησης στον κώδικα για την ευρύτερη κατηγορία crime_accident απεικονίζεται ως:

```
# Επιλογή των ειδήσεων που σχετίζονται με εγκλήματα-ατυχήματα
total_crime_accident_vocab = ['έγκλημα', 'ατύχημα', 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'αυτοκτονία',
                              'δυστύχημα', 'τροχαίο', 'κλοπή', 'διάρρηξη', 'βασανισμός', 'ξυλοδαρμός', 'βιασμός',
                              'γυναικοκτονία', 'παρανομία', 'λαθρεμπόριο', 'πορνεία', 'εμπρησμός', 'τρομοκρατία',
                              'επίθεση', 'σεξουαλική παρενόχληση', 'διακίνηση', 'κακοποίηση', 'εξαφάνιση',
                              'απαγωγή', 'bullying', 'εκφοβισμός', 'εκβιασμός', 'ασέλγεια', 'καραμπόλα', 'πνιγμός',
                              'αποπλάνηση', 'απάτη', 'απαγχονισμός', 'πυροβολισμός', 'βανδαλισμός', 'γενοκτονία',
                              'διαφθορά', 'ληστεία', 'δυσφήμιση', 'φοροδιαφυγή', 'ναρκωτικά', 'amber alert',
                              'πτώμα', 'νεκρός', 'τραυματίας', 'δίκη', 'δικαστήριο', 'ποινή', 'φυλακή',
                              'καταγγελία', 'πορνογραφία','δηλητηρίαση', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία',
                              'σωματεμπορία', 'πειρατεία', 'οπλοκατοχή', 'υπεξαίρεση', 'τραυματισμός', 'αεροπορικό',
                              'πετρελαιοκηλίδα', 'δουλεμπόριο', 'αρπαγή']
d = {ord('\N{COMBINING ACUTE ACCENT}'): None}  # αφαίρεση τόνων από τις ελληνικές λεξεις
total_crime_accident_df = categorization(total_crime_accident_vocab, title_totalvocab_stemmed, title_total_indexes,
                                         descr_totalvocab_stemmed, descr_total_indexes)
```

Εικόνα 45: Κλήση της συνάρτησης categorization στον κώδικα

Μετέπειτα απαραίτητη είναι η προσθήκη 2 στηλών στο DataFrame που ονομάζονται ¨sublevel¨ και ¨original_news_id¨. Συγκεκριμένα,

- **"total_crime_accident_df.insert(loc=6, column='sublevel', value='1')":** Αυτή η γραμμή εισάγει μια νέα στήλη με το όνομα "sublevel" στη θέση 6 στο DataFrame "total_crime_accident_df". Η παράμετρος `loc` καθορίζει τη θέση όπου πρέπει να εισαχθεί η νέα στήλη. Η παράμετρος «column» καθορίζει το όνομα της νέας στήλης, η οποία είναι «sublevel» σε αυτήν την περίπτωση. Η παράμετρος «value» καθορίζει την τιμή που πρέπει να εκχωρηθεί σε όλες τις εγγραφές στη νέα στήλη, η οποία είναι '1' σε αυτήν την περίπτωση, διότι η ευρύτερη κατηγορία crime_accident κωδικοποιείται με τον αριθμό '1', όπως φαίνεται και παραπάνω. Ουσιαστικά, αυτή η γραμμή προσθέτει μια νέα στήλη 'sublevel' στο DataFrame και τη γεμίζει με την τιμή '1' για όλες τις σειρές.
- **"total_crime_accident_df.insert(loc=7, column='original_news_id', value='0')":** Σημειώνεται πως η παρούσα γραμμή κώδικα επεξηγείται στην Ενότητα (8.2.2) της εργασίας [M23].

Συνοπτικά, αυτές οι δύο γραμμές κώδικα αυξάνουν το DataFrame «total_crime_accident_df» προσθέτοντας δύο νέες στήλες: «sublevel» και «original_news_id». Αυτές οι στήλες εισάγονται σε συγκεκριμένες θέσεις μέσα στο DataFrame και κάθε στήλη συμπληρώνεται με μια καθορισμένη τιμή για όλες τις σειρές.

Η συνάρτηση categorization γίνεται για όλες τις ευρύτερες κατηγορίες συμπεριλαμβανομένων των υποκατηγοριών τους. Η προσθήκη της στήλης sublevel γίνεται μόνο στις ευρύτερες κατηγορίες, όμως ο κωδικός κατηγορίας αλλάζει αν κάποια είδηση ανήκει σε υποκατηγορία της ευρύτερης κατηγορίας. Ο ρόλος και η διαχείριση της στήλης original_news_id επεξηγείται στην Ενότητα 8.2.2 της εργασίας [M23].

Στη συνέχεια εμφανίζεται ένα στιγμιότυπο από τον κώδικα όπου φαίνεται η εφαρμογή όλων των παραπάνω.

```
# Επιλογή των ειδήσεων που σχετίζονται με εγκλήματα-ατυχήματα
total_crime_accident_vocab = ['έγκλημα', 'ατύχημα', 'φόνος', 'δολοφονία', 'ανθρωποκτονία', 'αυτοκτονία',
                              'δυστύχημα', 'τροχαίο', 'κλοπή', 'διάρρηξη', 'βασανισμός', 'ξυλοδαρμός', 'βιασμός',
                              'γυναικοκτονία', 'παρανομία', 'λαθρεμπόριο', 'πορνεία', 'εμπρησμός', 'τρομοκρατία',
                              'επίθεση', 'σεξουαλική παρενόχληση', 'διακίνηση', 'κακοποίηση', 'εξαφάνιση',
                              'απαγωγή', 'bullying', 'εκφοβισμός', 'εκβιασμός', 'ασέλγεια', 'καραμπόλα', 'πνιγμός',
                              'αποπλάνηση', 'απάτη', 'απαγχονισμός', 'πυροβολισμός', 'βανδαλισμός', 'γενοκτονία',
                              'διαφθορά', 'ληστεία', 'δυσφήμιση', 'φοροδιαφυγή', 'ναρκωτικά', 'amber alert',
                              'πτώμα', 'νεκρός', 'τραυματίας', 'δίκη', 'δικαστήριο', 'ποινή', 'φυλακή',
                              'καταγγελία', 'πορνογραφία','δηλητηρίαση', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία',
                              'σωματεμπορία', 'πειρατεία', 'οπλοκατοχή', 'υπεξαίρεση', 'τραυματισμός', 'αεροπορικό',
                              'πετρελαιοκηλίδα', 'δουλεμπόριο', 'αρπαγή']
d = {ord('\N{COMBINING ACUTE ACCENT}'): None}  # αφαίρεση τόνων από τις ελληνικές λέξεις
total_crime_accident_df = categorization(total_crime_accident_vocab, title_totalvocab_stemmed, title_total_indexes,
                                         descr_totalvocab_stemmed, descr_total_indexes)
total_crime_accident_df.insert(loc=6, column='sublevel', value='1')
total_crime_accident_df.insert(loc=7, column='original_news_id', value='0')

if not total_crime_accident_df.empty:

    # Επιλογή των ειδήσεων που σχετίζονται με εγκλήματα
    total_crime_vocab = ['φόνος', 'δολοφονία', 'ανθρωποκτονία', 'κλοπή', 'διάρρηξη', 'βασανισμός', 'ξυλοδαρμός',
                         'βιασμός', 'γυναικοκτονία', 'παρανομία', 'λαθρεμπόριο', 'πορνεία', 'εμπρησμός',
                         'τρομοκρατία', 'απαγχονισμός', 'επίθεση', 'σεξουαλική παρενόχληση', 'διακίνηση', 'κακοποίηση',
                         'απαγωγή', 'bullying', 'εκφοβισμός', 'εκβιασμός', 'ασέλγεια', 'αποπλάνηση', 'απάτη',
                         'πυροβολισμός', 'βανδαλισμός', 'γενοκτονία', 'διαφθορά', 'ληστεία', 'δυσφήμιση', 'φοροδιαφυγή',
                         'ναρκωτικά', 'πορνογραφία', 'δηλητηρίαση', 'δωροληψία', 'δωροδοκία', 'πλαστογραφία',
                         'σωματεμπορία', 'πειρατεία', 'οπλοκατοχή', 'υπεξαίρεση', 'δουλεμπόριο', 'αρπαγή']
    d = {ord('\N{COMBINING ACUTE ACCENT}'): None}  # αφαίρεση τόνων από τις ελληνικές λέξεις
    total_crime_df = categorization(total_crime_vocab, title_totalvocab_stemmed, title_total_indexes,
                                    descr_totalvocab_stemmed, descr_total_indexes)

    sublevel(total_crime_accident_df, total_crime_df, '1.1')
```

Εικόνα 46: Παρουσίαση στιγμιότυπου κώδικα που σχετίζεται με την κατηγοριοποίηση

Επιπλέον, στο στιγμιότυπο εμφανίζεται μια συνθήκη που ουσιαστικά χρησιμοποιείται για την αξιοποίηση των πόρων του κώδικα. Αναλυτικότερα, επεξηγεί πως αν το DataFrame της ευρύτερης κατηγορίας ΔΕΝ είναι άδειο τότε να εισέλθει στο μπλοκ του κώδικα, αλλιώς να μην μπει στο μπλοκ κώδικα και να συνεχιστεί η εκτέλεση. Αυτό επίσης εφαρμόζεται και σε όλες τις κατηγορίες που βρίσκονται σε υψηλότερα επίπεδα και έχουν υποκατηγορίες.

8.14 Επιπλέον έλεγχος για την κατηγοριοποίηση

Μετά από την κατηγοριοποίηση, είναι αναγκαίος ένας περαιτέρω έλεγχος. Αυτός ο έλεγχος, έχει ως στόχο την απόρριψη ειδήσεων που παρόλο ανήκουν σε μια κατηγορία, στα πλαίσια της συγκεκριμένης πτυχιακής εργασίας, δεν θεωρούμε γεγονότα. Έτσι, δεν είναι επιθυμητή η εισαγωγή τους στην βάση δεδομένων και είναι απαραίτητη η αφαίρεσή τους από το DataFrame της ευρύτερης κατηγορίας όπου ανήκουν σύμφωνα με την συνάρτηση categorization.

Η κωδικοποίηση των κατηγοριών που έχει εκτελεστεί ο έλεγχος αυτός είναι οι **2, 7, 9, 10, 11**. Αυτό που ουσιαστικά συμβαίνει, είναι πως

- Πρώτα εισάγονται στο DataFrame της ευρύτερης κατηγορίας, οι προεπεξεργασμένες ειδήσεις, σύμφωνα με την συνάρτηση **categorization**
- Μετά, εκτελείται η συνάρτηση **title_description_filtering** στο αρχικό DataFrame που περιέχει όλες τις ειδήσεις και το αποτέλεσμα που θα προκύψει,
- Θα το βάλουμε ως όρισμα στην συνάρτηση **categorization** μαζί με το λεξικό που θα περιέχει λέξεις κλειδιά σχετικές με την κατηγορία αλλά αν βρεθούν σε κάποια είδηση, τότε θα την αφαιρέσουμε γιατί δεν την θεωρούμε γεγονός
- Τέλος, φιλτράρονται οι ειδήσεις χρησιμοποιώντας τη μέθοδο **.isin()** για να ελεγχθεί εάν οι τίτλοι των ειδήσεων στο DataFrame της κατηγορίας υπάρχουν στο DataFrame που περιέχει της ειδήσεις με τις λέξεις κλειδιά που δεν θεωρούμε γεγονότα. Το σύμβολο « ~ » χρησιμοποιείται ως λογικός τελεστής NOT για την αντιστροφή της συνθήκης. Ως αποτέλεσμα, καταργούνται αποτελεσματικά οι εγγραφές από το DataFrame της ευρύτερης κατηγορίας που έχουν τίτλους που ταιριάζουν με αυτούς του DataFrame που περιέχει ειδήσεις που δεν θεωρούνται γεγονότα.

Προκειμένου να γίνουν πιο κατανοητά τα παραπάνω βήματα, παρατίθεται ένα στιγμιότυπο από τον κώδικα από την κατηγορία περιβάλλον με την κωδικοποίηση «2».

Εικόνα 47: Επιπλέον έλεγχος κατηγοριοποίησης για την κατηγορία με κωδικό 2 (περιβάλλον)

Αυτό το τμήμα κώδικα ασχολείται με την επεξεργασία άρθρων ειδήσεων που σχετίζονται με το περιβάλλον. Αναλύεται παρακάτω ως:

1. Ορισμός λεξιλογίου περιβάλλοντος: Ορίζεται μια λίστα που ονομάζεται **total_environment_vocab**, η οποία περιέχει λέξεις κλειδιά που σχετίζονται με άρθρα ειδήσεων σχετικά με το περιβάλλον. Η μεταβλητή **d** ορίζεται ως λεξικό, το οποίο χρησιμοποιείται για την αφαίρεση τονισμού από ελληνικές λέξεις στο κείμενο. Αυτό, διασφαλίζει ότι οι λέξεις με διαφορετικούς τόνους αντιμετωπίζονται ως ίδιες.
2. Κατηγοριοποίηση άρθρων Περιβαλλοντικών Ειδήσεων: Αυτή η γραμμή χρησιμοποιεί τη λειτουργία κατηγοριοποίησης για την επεξεργασία και την κατηγοριοποίηση άρθρων ειδήσεων που σχετίζονται με το περιβάλλον. Λαμβάνει το λεξιλόγιο της κατηγορίας περιβάλλοντος, τις λέξεις του τίτλου προέλευσης, τους δείκτες του τίτλου, τις βασικές λέξεις περιγραφής και τους δείκτες της περιγραφής ως εισόδους. Το αποτέλεσμα αποθηκεύεται στο **total_environment_df** DataFrame (DataFrame ευρύτερης κατηγορίας), το οποίο περιέχει τα επεξεργασμένα και κατηγοριοποιημένα άρθρα ειδήσεων που σχετίζονται με το περιβάλλον.
3. Φιλτράρισμα άρθρων για την κατηγορία περιβάλλον: Σε αυτήν την ενότητα, εφαρμόζεται πρώτα η συνάρτηση **title_description_filtering** στο αρχικό DataFrame df για την επεξεργασία τίτλων και περιγραφών. Στη συνέχεια, ορίζεται ένα νέο περιβαλλοντικό λεξιλόγιο που περιέχει λέξεις όπως «καιρός» και «βροχή» οι οποίες δεν είναι αντιπροσωπευτικές για να θεωρηθεί μια είδηση γεγονός παρόλο που σχετίζονται με το περιβάλλον. Η συνάρτηση κατηγοριοποίησης χρησιμοποιείται ξανά για την επεξεργασία και την κατηγοριοποίηση άρθρων ειδήσεων με βάση αυτό το νέο λεξιλόγιο, αποθηκεύοντας το αποτέλεσμα στο **total_environment_df2**. Τέλος, φιλτράρει

τα ήδη επεξεργασμένα άρθρα που σχετίζονται με το περιβάλλον από το **total_environment_df** χρησιμοποιώντας τη μέθοδο **.isin()**.

4. Εισαγωγή στήλης: Εισάγονται δύο νέες στήλες, το **'sublevel'** και το **'original_news_id'**, στο total_environment_df DataFrame. Αυτό το βήμα εκχωρεί μια τιμή sublevel **'2'** για να υποδείξει την κατηγορία περιβάλλοντος και εκχωρεί το **'0'** ως το αρχικό αναγνωριστικό ειδήσεων για αυτά τα άρθρα σε κατηγορίες περιβάλλοντος σύμφωνα με την Ενότητα (8.2.2) στην εργασία [M23].

Συνοπτικά, αυτός ο κώδικας επεξεργάζεται άρθρα ειδήσεων που σχετίζονται με το περιβάλλον. Τα ταξινομεί, φιλτράρει τα διπλότυπα και εισάγει στήλες sublevel και original_news_id για να παρέχει πρόσθετες πληροφορίες σχετικά με τα άρθρα. Ο κώδικας παρουσιάζει μια συστηματική προσέγγιση για το χειρισμό των ειδήσεων που σχετίζονται με το περιβάλλον και την ενίσχυση της κατηγοριοποίησής τους.

Αντίστοιχα, αυτό συμβαίνει και με τις υπόλοιπες κατηγορίες και το μόνο που αλλάζει είναι τα ορίσματα και το λεξιλόγιο. Συγκεντρωτικά, τα λεξιλόγια για τις κατηγορίες είναι:

- Κωδικός κατηγορίας 2 (Περιβάλλον) → **‘καιρός’, ‘βροχή’**
- Κωδικός κατηγορίας 7 (Ολυμπιακοί αγώνες) → **‘μετάλλιο’, ‘νίκη’, ΄πρώτη’, ‘δεύτερη’, ‘τρίτη’, ‘θέση’, ‘διάκριση’**
- Κωδικός κατηγορίας 9 (Θρησκεία) → **‘γιορτή’, ‘εορτή’, ‘Πάσχα’, ‘Χριστούγεννα’**
- Κωδικός κατηγορίας 10 (Οικονομία) → **‘κέρδισε’**
- Κωδικός κατηγορίας 11 (Πολιτική) → **‘ομιλία’, ‘επίσκεψη’, ‘συνάντηση’**

8.15 Επίλογος

Συνοπτικά, στην συγκεκριμένη ενότητα (2.4), παρουσιάστηκαν οι λέξεις κλειδιά που χρησιμοποιούνται ως κριτήριο ύπαρξης για την κατάταξη ενός γεγονότος σε μια κατηγορία. Ακόμη αναλύθηκε ο τρόπος που γίνεται η κατηγοριοποίηση μειώνοντας όσο, είναι εφικτό, την πιθανότητα λάθους. Χρησιμοποιούνται διάφορες λειτουργίες και τεχνικές για την επεξεργασία πληροφοριών κειμένου, την αφαίρεση των διπλότυπων, τον υπολογισμό της ομοιότητας και την αποτελεσματική κατηγοριοποίηση των ειδήσεων. Βέβαια, επιβάλλεται να αναφερθεί και ο επιπλέον έλεγχος που χρειάζεται για την κατηγοριοποίηση προκειμένου να υπάρχει μεγαλύτερη

αξιοπιστία στα δεδομένα των ειδήσεων, με τεχνικές που αναλύονται στην πτυχιακή εργασία της Ρ. Μάντζαρη [M23].

9 Πειραματική μελέτη

9.1 Εισαγωγή

Για την πραγματοποίηση της πειραματικής μελέτης και κατ΄επέκταση την εξαγωγή στατιστικών προκειμένου να επαληθευτεί η λειτουργικότητα του κώδικα της συγκεκριμένης πτυχιακής εργασίας, έγινε συλλογή δεδομένων. Συγκεκριμένα, ο κώδικας εκτελέστηκε ασταμάτητα από την Παρασκευή 15 Σεπτεμβρίου 2023 (σύμφωνα με την κωδικοποίηση της συγκεκριμένης εργασίας 2023-09-15) στις 12:00 το πρωί έως την Πέμπτη 21 Σεπτεμβρίου 2023 (2023-09-21) στις 23:59. Στη συνέχεια παρουσιάζονται τα διαγράμματα για την καλύτερη κατανόηση των πραγμάτων, η εξέταση της ορθότητα του κώδικα και της λογικής που χρησιμοποιήθηκε.

9.2 Πλήθος γεγονότων ανά κατηγορία

Στον παρακάτω πίνακα απεικονίζεται το πλήθος των ειδήσεων ανά κατηγορία και το ποσοστό που καταλαμβάνει σε σχέση με τις συνολικές ειδήσεις.

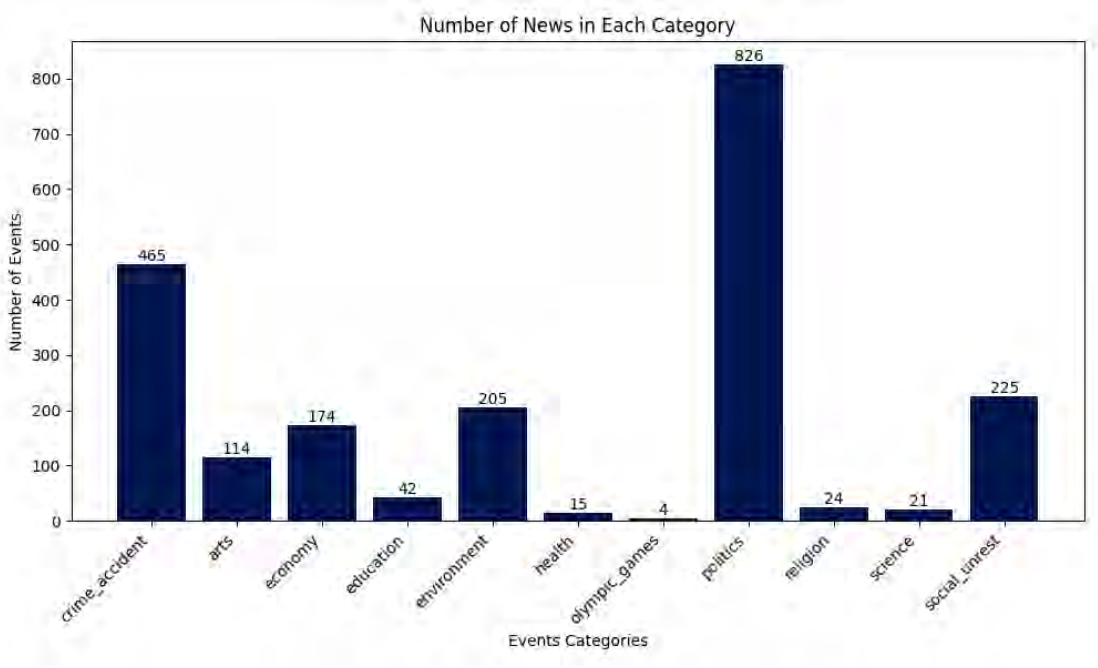
Πίνακας 2: Πλήθος-Ποσοστό γεγονότων ανά κατηγορία

Κατηγορία γεγονότων	Πλήθος γεγονότων	Ποσοστό
Crime_accident	465	21,9%
Arts	114	5,38%
Economy	174	8,22%
Education	42	1,98%
Environment	205	9,68%
Health	15	0,70%
Olympic_games	4	0,18%
Politics	826	39,03%
Religion	24	1,13%
Science	21	0,99%
Social_unrest	225	10,63%
Συνολικά	2116	~100%

Σύμφωνα με τον πίνακα, φαίνεται πως οι περισσότερες ειδήσεις βρίσκονται στην κατηγορία politics (πολιτική), λόγω των εκλογών του προέδρου του κόμματος

Σύριζα αλλά και στις δημοτικές και περιφερειακές εκλογές που θα διεξαχθούν τον μήνα Οκτώβριο.

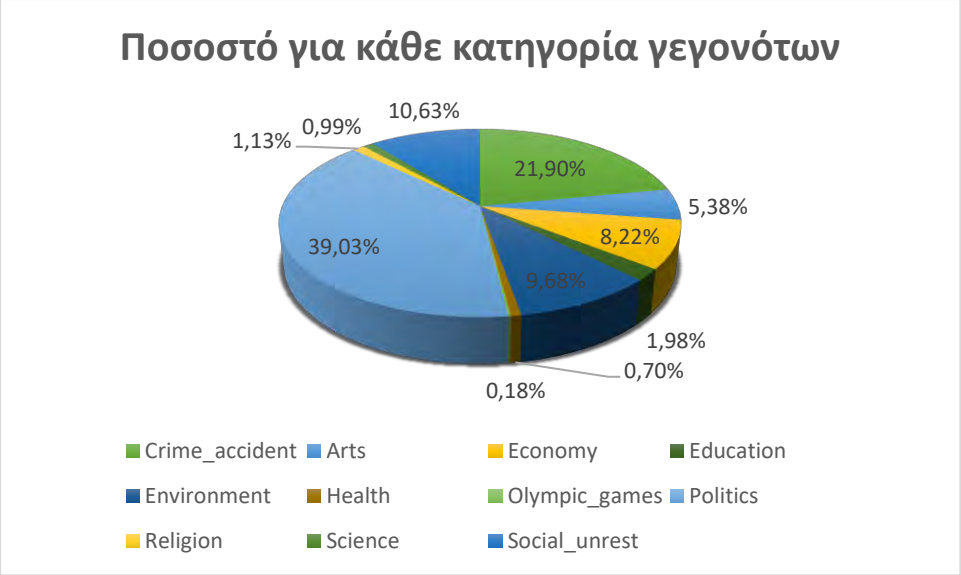
Στο διάγραμμα που ακολουθεί παρουσιάζονται τα αποτελέσματα του πίνακα σε γράφημα.



Εικόνα 48: Πλήθος γεγονότων ανά κατηγορία

Στην παραπάνω εικόνα παρουσιάζονται τα 11 Dataframes που αντιστοιχούν στις 11 κύριες κατηγορίες γεγονότων. Φαίνεται πως το dataframe των politics έχει τις περισσότερες ειδήσεις σε σχέση με τις υπόλοιπες κατηγορίες. Αυτό, οφείλεται και στην διεξαγωγή δημοτικών και περιφερειακών εκλογών τον μήνα Οκτώβριο αλλά και στην εκλογή νέου προέδρου του κόμματος του ΣΥΡΙΖΑ. Έπειτα, εμφανίζεται περισσότερο η κατηγορία crime_accident που στο μεγαλύτερο ποσοστό της αναφέρεται στο δυστύχημα της ελληνικής εθνικής άμυνας στην Λιβύη. Η αμέσως επόμενη κατηγορία είναι social_unrest αναφέρεται στις πορείες στην μνήμη του Παύλου Φύσσα και στην απεργία των ΜΜΜ. Με φθίνουσα σειρά ακολουθούν οι κατηγορίες environment, economy, arts, education, religion, science, health και τέλος olympic_games. Ο κώδικας που χρησιμοποιήθηκε για την δημιουργία του διαγράμματος μπάρας φαίνεται στο Παράρτημα-1.

Για την εμφάνιση του ποσοστού των ειδήσεων για κάθε κατηγορία, εμφανίζεται παρακάτω ένα διάγραμμα πίτας για την καλύτερη οπτικοποίηση και κατανόηση των αποτελεσμάτων που αναλύθηκαν με βάση την (Εικόνα 48).

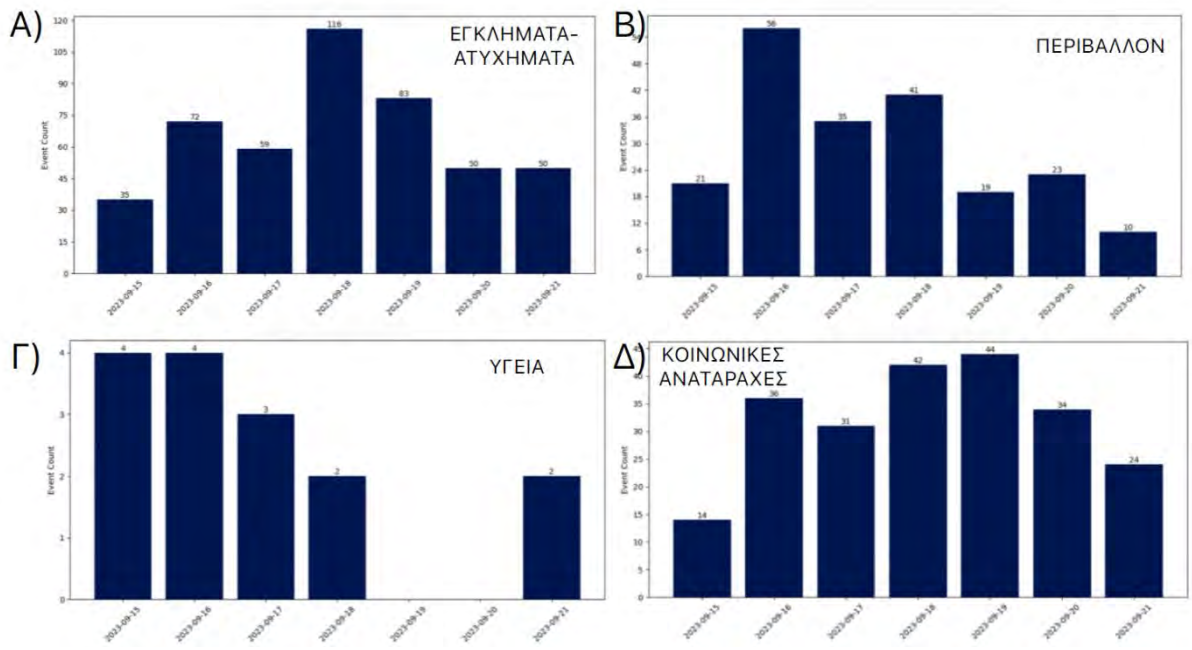


Εικόνα 49: Ποσοστό που καταλαμβάνει η κάθε κατηγορία σε σχέση με τα συνολικά γεγονότα.

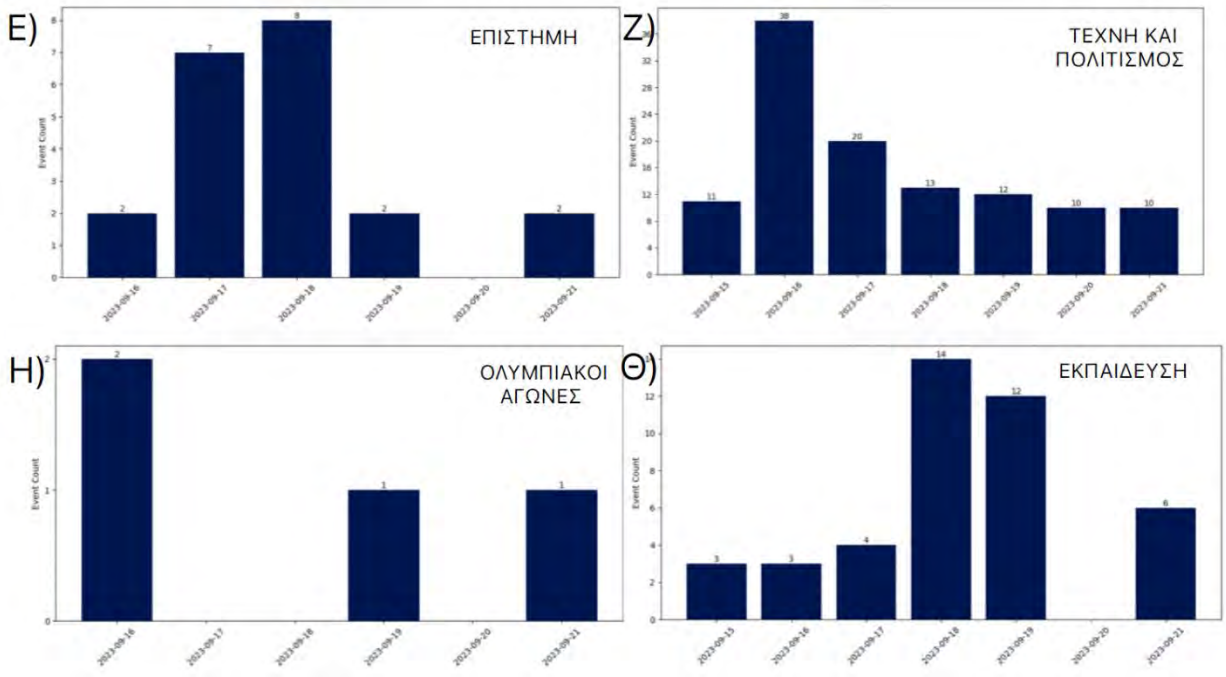
Και στο διάγραμμα πίτας φαίνεται φανερά πως η κατηγορία politics καταλαμβάνει το μεγαλύτερο μέρος της πίτας, στη συνέχεια είναι η κατηγορία crime_accident και έπειτα η κατηγορία social_unrest, ως οι κατηγορίες που καταλαμβάνουν το μεγαλύτερο ποσοστό συγκριτικά με τις υπόλοιπες.

9.3 Πλήθος γεγονότων ανά ημέρα και ανά κατηγορία

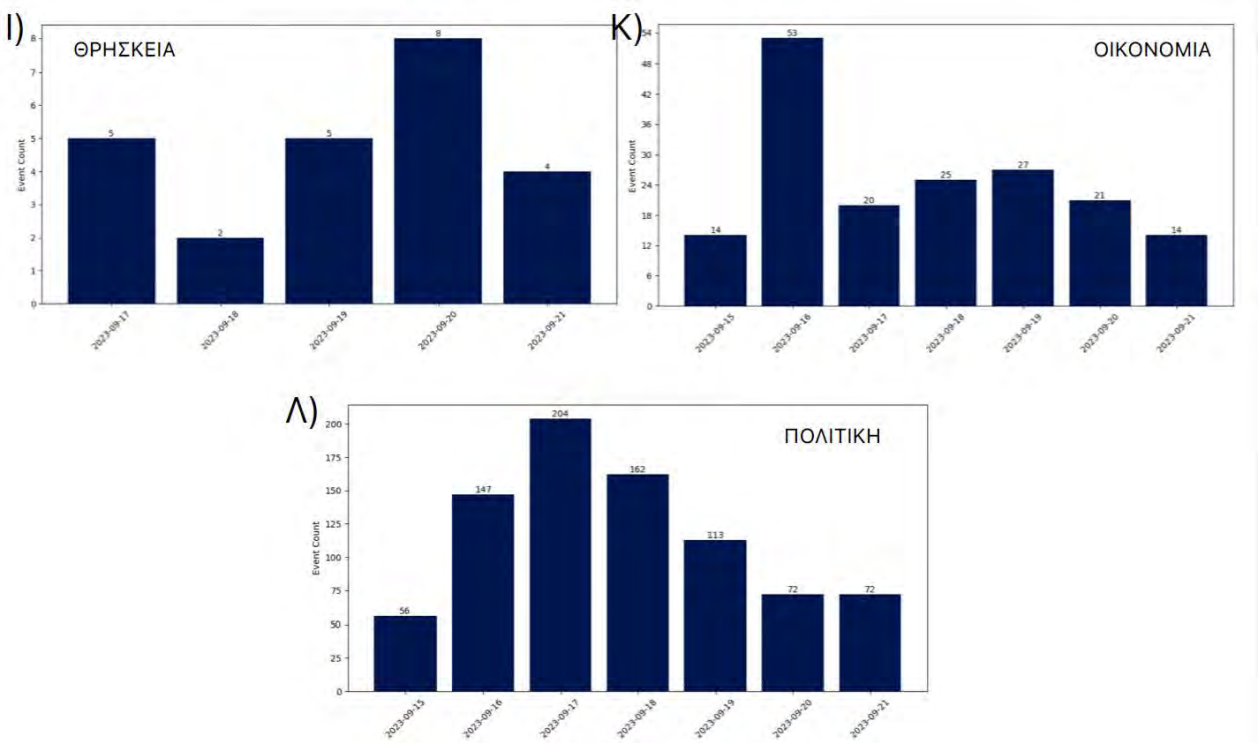
Στη συνέχεια αναφέρονται σε διαγράμματα το πλήθος των ειδήσεων που καταγράφηκαν αναλυτικά ανά ημέρα. Ο κώδικας που χρησιμοποιήθηκε για την δημιουργία των διαγραμμάτων φαίνεται στο Παράρτημα-2.



Εικόνα 50: Πλήθος γεγονότων ανά ημέρα για τις κατηγορίες Α) Εγκλήματα-Ατυχήματα, Β)Περιβάλλον, Γ)Υγεία, Δ)Κοινωνικές αναταραχές



Εικόνα 51: Πλήθος γεγονότων ανά ημέρα για τις κατηγορίες Ε)Επιστήμη, Ζ)Τέχνη και Πολιτισμός, Η)Ολυμπιακοί αγώνες, Θ)Εκπαίδευση



Εικόνα 52: Πλήθος γεγονότων ανά ημέρα για τις κατηγορίες Ι)Θρησκεία, Κ)Οικονομία, Λ)Πολιτική

Σε μερικά γραφήματα δεν απεικονίζονται ορισμένες ημέρες λόγω του ότι δεν υπάρχουν κάποια γεγονότα και αυτό συμβαίνει διότι βρίσκονται στις άκρες των διαγραμμάτων και άρα παραλείπονται, όπως στις κατηγορίες science και olympic_games. Όμως διαγράμματα όπως αυτό της κατηγορίας health, δεν υπάρχουν γεγονότα στις 2023-09-19 και 2023-09-20 και απεικονίζονται στο γράφημα επειδή βρίσκονται στην μέση του γραφήματος.

Τα συμπεράσματα που προκύπτουν από αυτή την υποενότητα είναι πως τις ημέρες Σάββατο 2023-09-16 και Δευτέρα 2023-09-18 εμφανίζεται η πλειονότητα των ειδήσεων στα περισσότερα διαγράμματα. Επίσης, εφόσον το πρόγραμμα ξεκίνησε από την Παρασκευή 2023-09-15 από τις 12 το μεσημέρι, μπορεί να εξηγεί το γεγονός ότι στα γραφήματα εκείνη την ημέρα δεν παρουσιάζονται τόσες ειδήσεις όσες αναμέναμε σε σχέση με τις υπόλοιπες μέρες.

9.4 Γράφημα Word Cloud ανά κατηγορία

Το σύννεφο λέξεων (word cloud) είναι ένα σύννεφο (cloud) γεμάτο με πολλές λέξεις σε διαφορετικά μεγέθη, που αντιπροσωπεύουν τη συχνότητα κάθε λέξης. Όσο πιο συχνά εμφανίζεται μια λέξη στο κείμενο, τόσο μεγαλύτερη θα εμφανίζεται στη

λέξη σύννεφο. Για την συγκεκριμένη υποενότητα, οι λέξεις που είναι πιο συχνές δείχνουν πως η πλειοψηφία των γεγονότων ανά κατηγορία αναφέρεται στις λέξεις αυτές [V23]. Ο κώδικας που χρησιμοποιήθηκε για την δημιουργία των word clouds φαίνεται στο Παράρτημα-3.

Εικόνα 53: Word cloud για τις κατηγορίες Α) Εγκλήματα-Ατυχήματα, Β)Περιβάλλον, Γ)Υγεία, Δ)Κοινωνικές αναταραχές

Εικόνα 54: Word cloud για τις κατηγορίες Ε)Επιστήμη, Ζ)Τέχνη και Πολιτισμός, Η)Ολυμπιακοί αγώνες, Θ)Εκπαίδευση

Εικόνα 55: Word cloud για τις κατηγορίες Ι)Θρησκεία, Κ)Οικονομία, Λ)Πολιτική

Σύμφωνα με τα παραπάνω διαγράμματα word cloud σε κάθε κατηγορία διακρίνονται οι πιο συχνά εμφανιζόμενες λέξεις. Συγκεκριμένα:

Α) *Εγκλήματα-Ατυχήματα (κωδικοποίηση κύριας κατηγορίας: 1):* Λιβύη, τροχαίο, δολοφονία. Αναφέρεται κυρίως στα θύματα που χάθηκαν στην Λιβύη από την κακοκαιρία Daniel, στα διάφορων ειδών τροχαία που συνέβησαν εκείνη την χρονική περίοδο αλλά και στην δολοφονία του Παύλου Φύσσα που έγινε πριν χρόνια.

Β) *Περιβάλλον (κωδικοποίηση κύριας κατηγορίας: 2):* Θεσσαλία, πλημμύρες, Λιβύη, φωτιά. Αναφέρεται κυρίως στις φονικές πλημμύρες που συνέβησαν στην Θεσσαλία και στην Λιβύη λόγω την κακοκαιρίας Daniel καθώς και στις φωτιές που έπληταν διάφορες περιοχές της Ελλάδας.

Γ) *Υγεία (κωδικοποίηση κύριας κατηγορίας: 3):* εμβόλιο, moderna, αντιγριπικό. Αναφέρεται κυρίως στο αντιγριπικό εμβόλιο που έφτιαξε η Moderna.

Δ) *Κοινωνικές αναταραχές (κωδικοποίηση κύριας κατηγορίας: 4):* απεργία, ΟΗΕ, μετανάστες. Αναφέρονται στην απεργία που έκαναν τα ΜΜΜ, την εμπλοκή του ΟΗΕ στα ζητήματα που αφορούν τους μετανάστες καθώς και στην Γενική Συνέλευση που πραγματοποιήθηκε.

Ε) *Επιστήμη (κωδικοποίηση κύριας κατηγορίας: 5):* moderna, αντιγριπικό, εμβόλιο. Αναφέρεται στην ανακάλυψη του νέου εμβολίου της Moderna ενάντια στη γρίπη.

Ζ) *Τέχνη και Πολιτισμός (κωδικοποίηση κύριας κατηγορίας: 6):* ΔΕΘ, Μητσοτάκης, Θεσσαλονίκη. Αναφέρεται κυρίως στην ΔΕΘ και την παρουσία του Μητσοτάκη σε αυτή.

Η) *Ολυμπιακοί αγώνες (κωδικοποίηση κύριας κατηγορίας: 7):* κορυφή, μεταλλίων, Ελλάδα. Αναφέρεται στο μετάλλιο που πήρε η Ελλάδα στους 3ους Μεσογειακούς Παράκτιους Αγώνες στο άθλημα της κολύμβησης.

Θ) *Εκπαίδευση (κωδικοποίηση κύριας κατηγορίας: 8):* Κασσελάκης, Θεσσαλία, τροχαίο. Αναφέρεται στις διαδικασίες που θα ακολουθήσει η Θεσσαλία όσον αφορά την εκπαίδευση, στο τροχαίο που συνέβη όπου το λεωφορείο είχε μαθητές ως επιβάτες και στις δηλώσεις του Κασσελάκη για την εκπαίδευση πριν γίνουν οι εκλογές για τον πρόεδρο του Σύριζα.

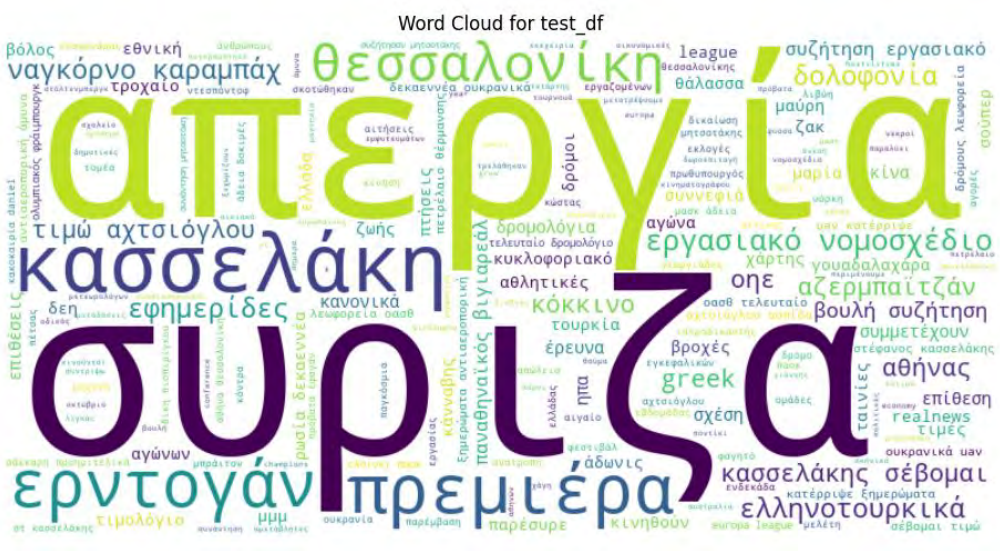
Ι) *Θρησκεία (κωδικοποίηση κύριας κατηγορίας: 9):* Λιβύη, σκοτώθηκαν, σύριζα. Αναφέρεται στα θύματα στην Λιβύη αλλά και στα ζητήματα που έχει το κόμμα Σύριζα σχετικά με την Τουρκία.

Κ) *Οικονομία (κωδικοποίηση κύριας κατηγορίας: 10):* Θεσσαλία, μέτρα, επιχειρήσεις. Αναφέρεται κυρίως στα οικονομικά μέτρα που θα παρθούν σχετικά με την κατάσταση στη Θεσσαλία λόγω των πλημμυρών.

Λ) *Πολιτική (κωδικοποίηση κύριας κατηγορίας: 11):* εκλογές σύριζα, Αχτσιόγλου, Μητσοτάκης. Αναφέρεται κυρίως στην εκλογή του προέδρου του κόμματος Σύριζα.

9.5 Γράφημα Word Cloud για την σύγκριση των μη φιλτραρισμένων με τις φιλτραρισμένες ειδήσεις

Για την πραγματοποίηση της σύγκρισης των μη φιλτραρισμένων σε σχέση με τις φιλτραρισμένες ειδήσεις χρησιμοποιήθηκε το word cloud προκειμένου να είναι εμφανής η διαφορά των ειδήσεων που υπάρχουν σύμφωνα με το περιεχόμενό τους. Η σύγκριση αυτή εκτελέστηκε για μία ημέρα και συγκεκριμένα για την Πέμπτη 21 Σεπτεμβρίου 2023 διότι αυτή την ημέρα υπήρχαν ειδήσεις που είχαν κατηγοριοποιηθεί σε όλες τις κατηγορίες. Ο κώδικας που χρησιμοποιήθηκε για την δημιουργία των word clouds είναι παρόμοιος με το Παράρτημα-3 και για τον λόγο αυτό παραλείπεται.



Εικόνα 56: Word cloud των μη φιλτραρισμένων ειδήσεων

Εικόνα 57: Word cloud των φιλτραρισμένων ειδήσεων-γεγονότων

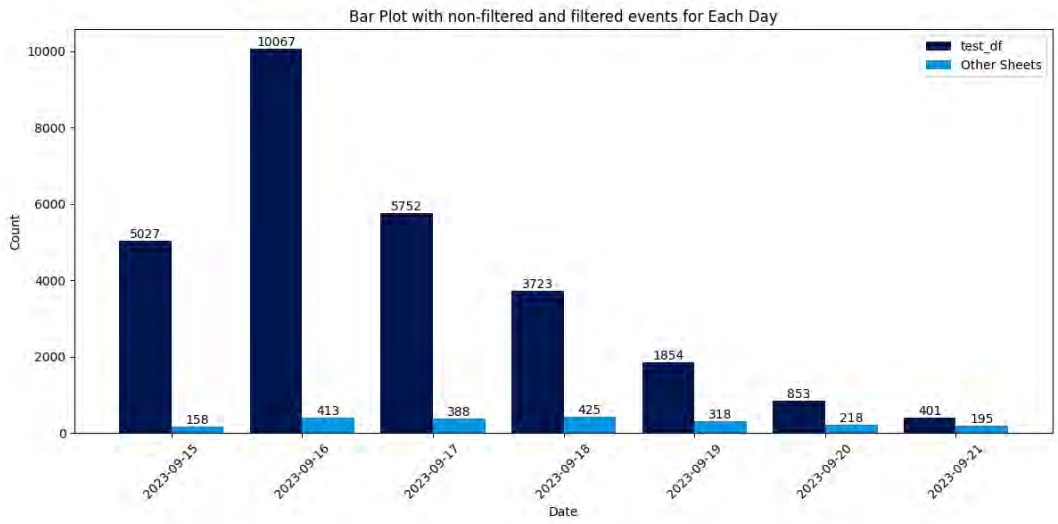
Στην εικόνα των μη φιλτραρισμένων ειδήσεων (πάνω εικόνα) γίνεται έμφαση στις λέξεις «συριζα», «απεργία» και «Κασσελάκη» ενώ στις φιλτραρισμένες ειδήσεις

(κάτω εικόνα) τονίζεται πολύ περισσότερο η λέξη «δολοφονία» και έπειτα οι λέξεις «συριζα» και «απεργία». Η διαφορά αυτή οφείλεται στο γεγονός των εκλογών του προέδρου του κόμματος Σύριζα την περίοδο που έγινε συλλογή των γεγονότων, με αποτέλεσμα να δίνεται έμφαση στις λέξεις «συριζα» και «Κασσελάκη» λόγω του μεγάλου αριθμού ειδήσεων για αυτή την θεματολογία. Βέβαια μετά το φιλτράρισμα η λέξη «δολοφονία» είναι πιο έντονη διότι υπάρχουν πάρα πολλά γεγονότα που επικεντρώνονται στην δολοφονία του Παύλου Φύσσα αλλά και η λέξη «απεργία» είναι έντονα τονισμένη λόγω τις κινητοποίησης των εργαζομένων για το νέο εργασιακό νομοσχέδιο.

Επιπλέον, μέσω της παρατήρησης των δυο word clouds εμφανίζονται στο word cloud των μη φιλτραρισμένων ειδήσεων και λέξεις όπως, «πρεμιέρα» και «league» όπου σχετίζονται κυρίως με το ποδόσφαιρο και δεν εμφανίζονται στο word cloud των φιλτραρισμένων ειδήσεων. Αυτό συμβαίνει λόγω της προεπεξεργασίας των raw data για την απόρριψη ειδήσεων που δεν θεωρούνται γεγονότα στα πλαίσια της παρούσας εργασίας.

9.6 Πλήθος φιλτραρισμένων και μη φιλτραρισμένων ειδήσεων ανά ημέρα

Στο παρακάτω γράφημα εμφανίζεται η διαφορά μεταξύ των φιλτραρισμένων σε σχέση με τις μη φιλτραρισμένες ειδήσεις. Ο κώδικας που χρησιμοποιήθηκε για την δημιουργία του διαγράμματος φαίνεται στο Παράρτημα-4.



Εικόνα 58: Πλήθος φιλτραρισμένων και μη φιλτραρισμένων ειδήσεων ανά ημέρα

Τα αποτελέσματα που προκύπτουν από το συγκεκριμένο γράφημα είναι πως εμφανίζονται με μεγάλη διαφορά το πλήθος των ειδήσεων που μαζεύει ο κώδικας

συγκριτικά με τις ειδήσεις που κατηγοριοποιεί. Αυτό συμβαίνει λόγω της προεπεξεργασίας που έχει πραγματοποιηθεί προκειμένου να κατηγοριοποιούνται μόνο οι ειδήσεις που θεωρούνται γεγονότα σύμφωνα με τις προδιαγραφές που παρουσιάζονται στην ενότητα 1.1.3 από το μέρος Α, του συγκεκριμένου εγγράφου, ενώ οι υπόλοιπες ειδήσεις δεν θεωρούνται γεγονότα και απορρίπτονται.

Την ημέρα 2023-09-16 παρουσιάζεται η μεγαλύτερη αλλαγή μεταξύ φιλτραρισμένων και μη φιλτραρισμένων ειδήσεων. Κάτι τέτοιό οφείλεται σε μεγάλο πλήθος ειδήσεων που δεν θεωρούνται γεγονότα. Μερικά παραδείγματα τέτοιων περιπτώσεων είναι ειδήσεις σχετικά με Έλληνες και ξένους celebrities, ειδήσεις σχετικά με τι καιρό θα κάνει σε διάφορα μέρη στην Ελλάδα, ειδήσεις σχετικά με διατροφικές και ιατρικές συμβουλές, ειδήσεις με διαγωνισμούς δώρων, ειδήσεις σε αθλητικές εφημερίδες, αγγλικές ειδήσεις καθώς και ειδήσεις σχετικές με ονομαστικές γιορτές.

Σε αντίθεση, βέβαια, την ημέρα 2023-09-21 παρατηρείται η μικρότερη διαφορά μεταξύ φιλτραρισμένων και αφιλτράριστων ειδήσεων. Η περίπτωση αυτή οφείλεται σε μικρότερης συχνότητας ειδήσεις που σχετίζονται με το lifestyle, αθλητικά νέα και διαγωνισμούς δώρων στους πολίτες, τα οποία δεν θεωρούνται γεγονότα.

Επιπλέον, ο κώδικας συλλέγει τις ειδήσεις ανά μία ώρα, επομένως το feed κάθε ειδησεογραφικού δεν είναι πλήρως ανανεωμένο με νέες ειδήσεις με αποτέλεσμα να συλλέγεται η ίδια είδηση πολλές φορές.

9.7 Υπολογισμός μετρικών για κατηγοριοποίηση

Για την επαλήθευση της διαδικασίας του πληροφοριακού συστήματος, πραγματοποιήθηκε χειροκίνητα η κατηγοριοποίηση των γεγονότων που συλλέχθηκαν στην ημέρα Πέμπτη 2023-09-21. Δεν ήταν εφικτή η κατηγοριοποίηση όλων των ειδήσεων που συλλέχθηκαν το χρονικό διάστημα 15-21 Σεπτεμβρίου 2023 λόγω του μεγάλου πλήθους τους, συγκεκριμένα 27.678 ειδήσεις. Επιλέχθηκε, λοιπόν, για τον λόγο αυτό, η ημέρα 2023-09-21 επειδή βρισκόντουσαν ειδήσεις κατηγοριοποιημένες σε όλες τις κατηγορίες γεγονότων προκειμένου να παρθεί το αποτέλεσμα από ένα αντιπροσωπευτικό δείγμα γεγονότων και συνολικά ήταν 401. Το dataset που χρησιμοποιήθηκε για τον συγκεκριμένο έλεγχο βρίσκεται στο αρχείο Παραδοτέο-2_metrics_for_categorization.

Αφού έγινε η συλλογή των ειδήσεων, διαπιστώθηκε ποιες ειδήσεις θα χαρακτηριστούν ως γεγονότα και ποιες ως μη γεγονότα, προκειμένου στη συνέχεια να κατηγοριοποιηθούν. Για τις αποφάσεις αυτές βοήθησε η συνάρτηση Κατηγοριοποίηση(Χ), όπου Χ είναι μια είδηση. Για την αποδοτικότητα της συνάρτησης αυτής έγινε καταμέτρηση των TP, TN, FP και FN. Ο έλεγχος της κατηγοριοποίησης πραγματοποιήθηκε σύμφωνα με τα παρακάτω κριτήρια:

1. Έστω ότι για έναν άνθρωπο-παρατηρητή η είδηση Α περιγράφει ένα πολιτικό γεγονός. Κατέταξε η συνάρτηση Κατηγοριοποίηση(Α) την είδηση Α στα πολιτικά γεγονότα; Αν ναι, τότε έχουμε TP. Αν την κατέταξε σε λάθος κατηγορία (π.χ. στο αντικείμενο Υγεία) ή αν δεν την κατέταξε καθόλου στο δένδρο, τότε έχουμε FN. Αν την κατέταξε σε δύο λάθος κατηγορίες (π.χ. στο αντικείμενο Υγεία και στο αντικείμενο Περιβάλλον), τότε έχουμε δύο FN. Αν την κατέταξε σε δύο κατηγορίες από τις οποίες η μία είναι σωστή (π.χ. στο αντικείμενο Πολιτική και στο αντικείμενο Υγεία), τότε έχουμε ένα TP και ένα FN.
2. Για τον ίδιο άνθρωπο-παρατηρητή η είδηση Β περιγράφει ένα πολιτικό γεγονός. Κατέταξε η συνάρτηση Κατηγοριοποίηση(Β) την είδηση Β στα πολιτικά γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε TP ενώ αν είναι αρνητική, τότε έχουμε FN.
3. Έστω ότι για τον άνθρωπο-παρατηρητή η είδηση Γ δεν περιγράφει γεγονός. Κατέταξε η συνάρτηση την είδηση Γ στα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε FP ενώ αν είναι αρνητική, τότε έχουμε TN.
4. Έστω ότι για τον άνθρωπο-παρατηρητή η είδηση Δ δεν περιγράφει γεγονός. Κατέταξε η συνάρτηση την είδηση Δ στα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε FP ενώ αν είναι αρνητική, τότε έχουμε TN.
5. Τέλος η είδηση Ε. Έστω ότι για τον άνθρωπο-παρατηρητή η είδηση αυτή περιγράφει ένα γεγονός πολιτικό και κοινωνικής αναταραχής (δύο κατηγορίες). Κατέταξε η συνάρτηση την είδηση Ε και στις δύο σωστές κατηγορίες γεγονότων; Αν η απάντηση είναι θετική, τότε έχουμε δύο TP ενώ αν είναι αρνητική, τότε έχουμε δύο FN ή ίσως ένα TP και ένα FN, ανάλογα με την αποδοτικότητα της συνάρτησης. Δηλαδή αν η είδηση αποθηκεύτηκε μόνο στην κατηγορία Επιστήμη, τότε έχουμε δύο FN διότι η είδηση Ε δεν προκάλεσε στη

βάση δεδομένων τις δύο εισαγωγές στις κατηγορίες Πολιτική και Κοινωνικές αναταραχές, όπως περιμέναμε.

Ακολουθώντας πιστά τα κριτήρια που αναφέρθηκαν, τα αποτελέσματα που προέκυψαν φαίνονται παρακάτω:

- TP = 161
- TN = 200
- FP = 9
- FN = 31

Εύλογα, λοιπόν, μπορούν να προκύψουν αποτελέσματα για τις μετρικές. Αναλυτικότερα:

- Ορθότητα (Accuracy) = (TP+TN)/(TP+TN+FP+FN) = 0,90

Το 90% των γεγονότων κατηγοριοποιήθηκε σωστά.

- Ανάκληση (Recall) = TP/(TP+FN) = 0,84

Το 0,84 υποδηλώνει πως το μοντέλο καταγράφει κατά 84% τις πραγματικές θετικές περιπτώσεις, ελαχιστοποιώντας τον αριθμό των ψευδώς αρνητικών.

- Ακρίβεια (Precision) = TP/(TP+FP) = 0,94

Το 94% της κατηγοριοποίησης κάνει θετικές προβλέψεις που είναι πραγματικά θετικές.

- F1-score = 2*TP/(2*TP + FP + FN) = 0,89

Το F1-Score είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης. Παρέχει μια ισορροπία μεταξύ ακρίβειας και ανάκλησης, ιδίως όταν υπάρχει ανισορροπία μεταξύ των κλάσεων. Το F1-Score 0,89 υποδηλώνει μια καλή ισορροπία μεταξύ της πραγματοποίησης ακριβών θετικών προβλέψεων και της σύλληψης των περισσότερων πραγματικών θετικών αποτελεσμάτων.

Συμπερασματικά, υπάρχει υψηλή ορθότητα, ακρίβεια και F1-Score, υποδεικνύοντας πως η κατηγοριοποίηση των γεγονότων έχει καλή απόδοση όσον αφορά την πραγματοποίηση ορθών θετικών προβλέψεων. Η ανάκληση είναι ελαφρώς χαμηλότερη από την ακρίβεια, γεγονός που υποδηλώνει ότι η κατηγοριοποίηση μπορεί

να είναι πιο προσεκτική στην πραγματοποίηση θετικών προβλέψεων, αλλά όταν κάνει μια θετική πρόβλεψη, είναι συχνά σωστή.

10 Συμπεράσματα και προτάσεις για μελλοντική εργασία

10.1 Συμπεράσματα

Εν κατακλείδι, στην παρούσα πτυχιακή εργασία, παρουσιάστηκε το πρώτο μεγάλος μέρος από την ευρεία έρευνα που είχαμε στόχο με την συνερευνήτριά μου Ρ. Μάντζαρη, στο οποίο εκείνη συνέχισε με την υλοποίηση του δεύτερου μέρους [M23]. Αναλυτικότερα, στην παρούσα πτυχιακή εργασία περιγράφηκε πώς επιτεύχθηκε ο στόχος της συλλογής, αναγνώρισης και κατηγοριοποίησης των γεγονότων σε πραγματικό χρόνο και η αυτοματοποιημένη αποθήκευσή τους μέσα σε βάση δεδομένων. Με τον τρόπο αυτό, συγκεντρώθηκαν ποιοτικά δεδομένα καθώς έγινε ο διαχωρισμός των ειδήσεων από τα γεγονότα αλλά και η οργάνωση των πιο σημαντικών πληροφοριών από τα γεγονότα στα πλαίσια της πτυχιακής μου. Τα επόμενα βήματα, τα οποία περιγράφονται στην εργασία [M23], είναι η ανίχνευση επανεκδόσεων ειδησεογραφικών άρθρων, βασισμένη στα δεδομένα που προέκυψαν ως τελικό προϊόν της δικής μου υλοποίησης, και στη συνέχεια η εύρεση ειδήσεων που περιγράφουν την εξέλιξη των γεγονότων με την πάροδο του χρόνου.

Σύμφωνα με την πειραματική μελέτη που παρουσιάστηκε στην παρούσα πτυχιακή εργασία, φαίνεται πως το αποτέλεσμα αυτής της υλοποίησης είναι αξιοσημείωτο καθώς το σύστημα λειτουργεί ορθά με ποσοστό 90%. Αυτό φυσικά δεν αναιρεί το γεγονός πως με τις κατάλληλες διορθώσεις το μπορεί να λειτουργήσει και ακόμη καλύτερα, γι'αυτό στις επόμενες υποενότητες παρατίθενται προτάσεις για μελλοντική εργασία σχετικά με τη βελτιστοποίηση της λειτουργίας του και των υπηρεσιών που μπορεί να προσφέρει.

10.2 Προτάσεις για μελλοντική εργασία: Δημιουργία ιστοσελίδας για την εύκολη πρόσβαση των πληροφοριών της βάσης δεδομένων

Στην επιδίωξη της βελτίωσης της συλλογής ειδήσεων και της προσβασιμότητας, η δημιουργία ενός αποκλειστικού ιστότοπου συνδεδεμένου με τη βάση δεδομένων ειδήσεων αποτελεί έναν συνδετικό κρίκο για τη βελτίωση της αλληλεπίδρασης των χρηστών και της ανάκτησης πληροφοριών. Σχεδιάζοντας μια φιλική προς τον χρήστη διαδικτυακή πλατφόρμα που ενσωματώνεται απρόσκοπτα στη βάση δεδομένων, οι χρήστες μπορούν να έχουν πρόσβαση, να αναζητούν και να εξερευνούν αβίαστα άρθρα ειδήσεων.

Ένας αποκλειστικός ιστότοπος παρέχει μια οικεία και διαισθητική διεπαφή για τους χρήστες να αλληλεπιδρούν με τη βάση δεδομένων ειδήσεων. Μέσω μιας εύκολης πλοήγησης διάταξης και διαδραστικών λειτουργιών, οι χρήστες μπορούν να αναζητούν, να φιλτράρουν και να ανακτούν γρήγορα άρθρα ειδήσεων με βάση τις προτιμήσεις και τα ενδιαφέροντά τους.

Οι προηγμένες επιλογές αναζήτησης, όπως η αναζήτηση λέξεων-κλειδιών, η επιλογή εύρους ημερομηνιών και το φιλτράρισμα κατηγοριών, επιτρέπουν στους χρήστες να ανακαλύπτουν αποτελεσματικά τις επόμενες ειδήσεις που σχετίζονται με τα ερωτήματά τους. Επιπλέον, η ενσωμάτωση ετικετών που προέρχονται από το ονομαζόμενο σύστημα αναγνώρισης οντοτήτων μπορεί να βελτιώσει περαιτέρω τα αποτελέσματα αναζήτησης και να διευκολύνει την ακριβή ανάκτηση πληροφοριών.

Ο ιστότοπος μπορεί να αξιοποιήσει τεχνικές οπτικοποίησης δεδομένων για να παρουσιάσει τάσεις, μοτίβα και ιδέες που προέρχονται από τη βάση δεδομένων ειδήσεων. Τα διαδραστικά γραφήματα, τα γραφήματα και οι χάρτες θερμότητας μπορούν να παρέχουν στους χρήστες μια οπτική κατανόηση της κάλυψης ειδήσεων με την πάροδο του χρόνου, των διαδεδομένων θεμάτων και των σχετικών οντοτήτων. Αυτό όχι μόνο ενισχύει την αφοσίωση των χρηστών, αλλά προσφέρει επίσης μια ολοκληρωμένη προοπτική για τις εξελίξεις των ειδήσεων.

Καθώς η βάση δεδομένων ειδήσεων μεγαλώνει με την πάροδο του χρόνου, ο ιστότοπος θα πρέπει να σχεδιαστεί για να χειρίζεται αυξανόμενους όγκους δεδομένων και επισκεψιμότητας χρηστών. Η κλιμακωτή αρχιτεκτονική, η αποτελεσματική διαχείριση της βάσης δεδομένων και η τακτική συντήρηση είναι απαραίτητα για τη διασφάλιση της ομαλής απόδοσης της πλατφόρμας και την πρόσβαση σε άρθρα ειδήσεων.

10.2.1 Ενσωμάτωση λήψεων φύλλων του Excel από την ιστοσελίδα

Εκτός από την παροχή μιας διαισθητικής διεπαφής για ανακάλυψη ειδήσεων, η ενσωμάτωση της δυνατότητας λήψης φύλλων του Excel ή αρχείων csv απευθείας από τον ιστότοπο μπορεί να βελτιώσει περαιτέρω την αφοσίωση των χρηστών και την προσβασιμότητα των δεδομένων. Τα φύλλα του Excel χρησιμεύουν ως μια ευρέως αναγνωρισμένη μορφή για την οργάνωση και την ανάλυση δεδομένων.

Επιτρέποντας στους χρήστες να κατεβάζουν δεδομένα ειδήσεων σε μορφή Excel ή csv, τους επιτρέπει να εξάγουν δομημένες πληροφορίες, όπως τίτλους,

ημερομηνίες δημοσίευσης, συγγραφείς, κατηγορίες και επώνυμες οντότητες. Αυτό διευκολύνει την ολοκληρωμένη ανάλυση και τις ιδέες που μπορούν να βοηθήσουν τους ερευνητές, τους δημοσιογράφους και τους αναλυτές να κατανοήσουν τις τάσεις και τις εξελίξεις των ειδήσεων.

Οι ερευνητές και οι αναλυτές απαιτούν συχνά ακατέργαστα δεδομένα για εις βάθος μελέτη και εξερεύνηση. Παρέχοντας φύλλα Excel ή αρχεία csv με δυνατότητα λήψης, ο ιστότοπος απευθύνεται σε χρήστες που επιδιώκουν να πραγματοποιήσουν ανεξάρτητη έρευνα και να πραγματοποιήσουν ανάλυση βάσει δεδομένων. Η ενσωμάτωση της δυνατότητας λήψης φύλλων του Excel ή αρχείων csv από τον ιστότοπο προσθέτει ένα επίπεδο ευελιξίας στην εμπειρία του χρήστη, εξυπηρετώντας ερευνητές και αναλυτές που αναζητούν δομημένα δεδομένα για ανάλυση.

10.3 Χρήση Βαθιάς Μάθησης για την περαιτέρω αυτοματοποίηση του πληροφοριακού συστήματος

Η βαθιά μάθηση μπορεί να διαδραματίσει σημαντικό ρόλο στην αυτοματοποίηση ενός πληροφοριακού συστήματος που χρησιμοποιείται για τη συλλογή ειδήσεων. Η περίπλοκη και εξελισσόμενη φύση των δεδομένων ειδήσεων απαιτεί προηγμένες τεχνικές για την αποτελεσματική εξαγωγή, κατηγοριοποίηση και ανάλυση πληροφοριών.

1. Ταξινόμηση και κατηγοριοποίηση κειμένων: Μπορούν να χρησιμοποιηθούν μοντέλα βαθιάς μάθησης, όπως Συνελικτικά νευρωνικά δίκτυα (CNN) ή αναδρομικά νευρωνικά δίκτυα (RNN), για την αυτόματη κατηγοριοποίηση των άρθρων ειδήσεων σε διάφορες κατηγορίες. Αυτά τα μοντέλα μπορούν να μάθουν να αναγνωρίζουν μοτίβα και χαρακτηριστικά στο κείμενο, επιτρέποντας την ακριβή ταξινόμηση των άρθρων με βάση το περιεχόμενό τους.
2. Καθαρισμός και προεπεξεργασία δεδομένων: Η βαθιά μάθηση μπορεί να βοηθήσει στον καθαρισμό και την προεπεξεργασία ακατέργαστων δεδομένων ειδήσεων. Τα μοντέλα μπορούν να διορθώνουν ορθογραφικά λάθη, να αφαιρούν θόρυβο και να τυποποιούν το κείμενο, διασφαλίζοντας ότι τα δεδομένα που συγκεντρώνονται είναι υψηλής ποιότητας.

Συνολικά, η βαθιά μάθηση μπορεί να αυτοματοποιήσει σημαντικά τη διαδικασία συλλογής, ανάλυσης και οργάνωσης ειδήσεων αξιοποιώντας την ικανότητα

των μοντέλων να μαθαίνουν περίπλοκα μοτίβα και αναπαραστάσεις από μεγάλες ποσότητες δεδομένων κειμένου. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η ανάπτυξη και η τελειοποίηση αυτών των μοντέλων απαιτεί σημαντικά δεδομένα με ετικέτα, υπολογιστικούς πόρους και εξειδίκευση σε τεχνικές βαθιάς μάθησης.

11 Παραρτήματα

Παράρτημα-1

```python
# Number of news by category
import pandas as pd
import matplotlib.pyplot as plt

file_path = 'total_duplicates_categorization.xlsx'
xl = pd.ExcelFile(file_path)

# Create an empty list to store the results
results = []

# Get the sheet names from the Excel file
sheet_names = xl.sheet_names

# Iterate through the sheets and count the rows in each sheet
for sheet_name in sheet_names:
    if sheet_name == 'test_df':
        continue
    # Read each sheet into a DataFrame
    df = xl.parse(sheet_name)

    # Get the number of rows in the DataFrame
    num_rows = len(df)

    # Append the results to the list
    results.append({'Sheet Name': sheet_name, 'Number of Rows': num_rows})

# Create a DataFrame from the results list
results_df = pd.DataFrame(results)

# Now you have a DataFrame with sheet names and their corresponding row counts
print(results_df)

# Create a bar plot using matplotlib
plt.figure(figsize=(10, 6))
bars = plt.bar(results_df['Sheet Name'], results_df['Number of Rows'], color='#01154E')

# Add annotations (number of rows) above each bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval, int(yval), ha='center', va='bottom')

plt.xlabel('Events Categories')
plt.ylabel('Number of Events')
plt.title('Number of News in Each Category')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

# Save the plot to a file if needed
plt.savefig(f'total_events_barplot.png')

plt.show()
```

Εικόνα 59: Κώδικας για την υλοποίηση του διαγράμματος σχετικά με το πλήθος γεγονότων ανά κατηγορία.

Παράρτημα-2

```python
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.ticker import MaxNLocator

# Load the Excel file
file_path = "total_duplicates_categorization.xlsx"
xls = pd.ExcelFile(file_path)

# Get the sheet names (assuming each sheet corresponds to a category)
sheet_names = xls.sheet_names

# Define the date format
date_format = '%Y-%m-%d %H:%M:%S'

# Initialize an empty DataFrame to store aggregated data
all_data = pd.DataFrame(columns=['pubdate', 'category'])
data = {}

# Read data from each sheet, convert pubDate to datetime, and concatenate them
for sheet_name in sheet_names:
    if sheet_name == 'test_df':
        continue
    df = pd.read_excel(file_path, sheet_name=sheet_name)
    df['pubdate'] = pd.to_datetime(df['pubdate'], format=date_format)
    df['pubdate'] = df['pubdate'].dt.strftime('%Y-%m-%d')  # Format date
    data[sheet_name] = df
    # Convert the date column to datetime format
    df['pubdate'] = pd.to_datetime(df['pubdate'])

    # Group the DataFrame by date and count the rows for each date
    date_counts = df.groupby(df['pubdate'].dt.date).size().reset_index(name='count')
    print(date_counts)

    ax = plt.figure(figsize=(10, 6)).gca()  # Adjust the figure size as needed
    ax.yaxis.set_major_locator(MaxNLocator(integer = True))
    bar_name = plt.bar(date_counts['pubdate'], date_counts['count'], color='#01154E')


    # Annotate each bar with the row count directly under the bar
    for i, count in enumerate(date_counts['count']):
        plt.text(date_counts['pubdate'][i], count, str(count), ha='center', va='bottom')

    plt.title(f'Daily Event Counts for {sheet_name}')
    plt.xlabel('Date')
    plt.ylabel('Event Count')
    date_range = pd.date_range(start=date_counts['pubdate'].iloc[0], end=date_counts['pubdate'].iloc[-1], freq='D')
    date_range = date_range.strftime('%Y-%m-%d')
    plt.xticks(date_range,rotation=45)  # Rotate x-axis labels for better visibility
    plt.tight_layout()

    # Save the plot to a file if needed
    plt.savefig(f'{sheet_name}_barplot.png')

    # Show the plot
    plt.show()
print(data)
```

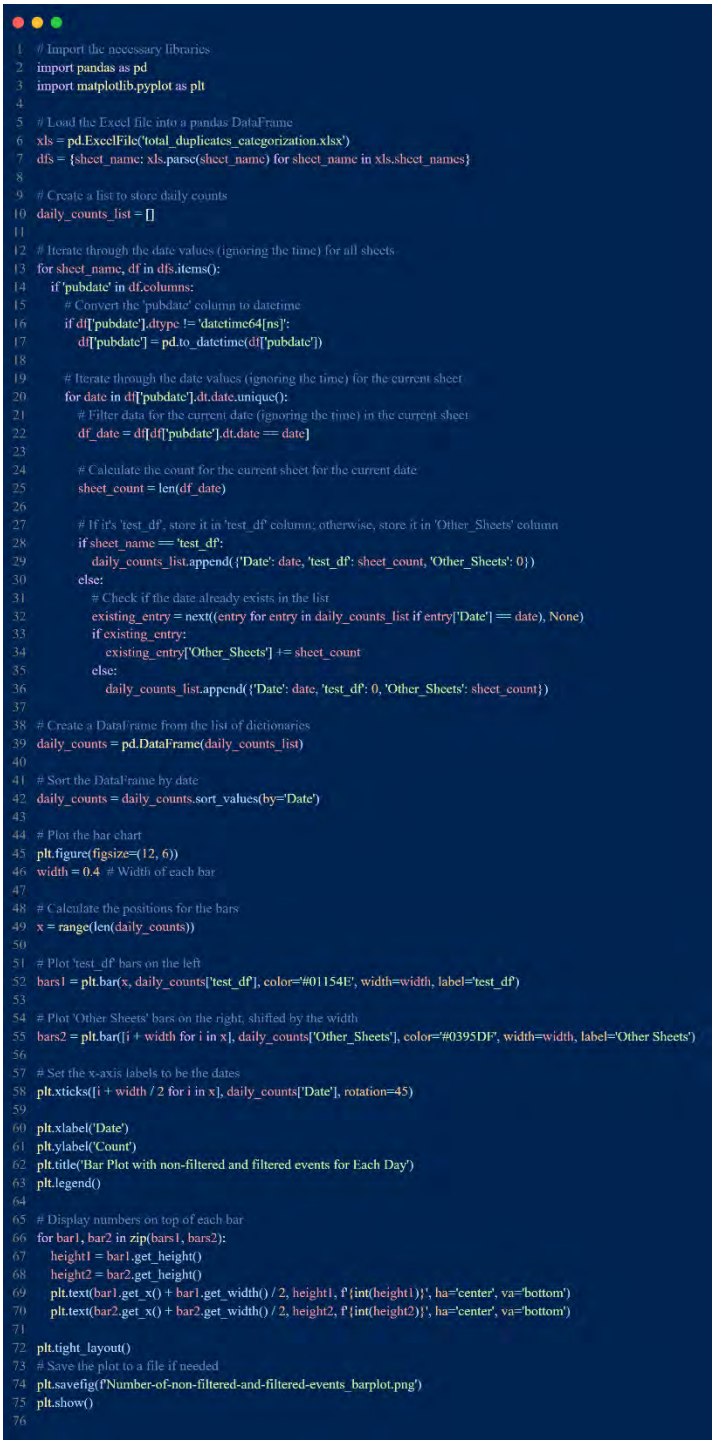
Εικόνα 60: Κώδικας για την υλοποίηση των διαγραμμάτων σχετικά με το πλήθος γεγονότων ανά ημέρα για κάθε κατηγορία ξεχωριστά.

Παράρτημα-3

```python
# word cloud
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import nltk
import os

# Download NLTK stopwords data for Greek
# nltk.download('stopwords')
# nltk.download('punkt')

# Load the Excel file
excel_file = 'total_duplicates_categorization.xlsx'
xlsx = pd.ExcelFile(excel_file)

# Load custom Greek stop words from 'extendstopwords.txt'
stop_words_file = 'extendstopwords.txt'  # Make sure this file contains Greek stop words
with open(stop_words_file, 'r', encoding='utf-8') as f:
    greek_stopwords = set(f.read().splitlines())

# Create a directory to save the word cloud images
output_dir = 'wordclouds_duplicates_only'
os.makedirs(output_dir, exist_ok=True)

# Process each sheet
for sheet_name in xlsx.sheet_names:

    # Read the sheet into a DataFrame
    df = pd.read_excel(excel_file, sheet_name=sheet_name)

    # Combine all titles into a single string
    titles = ' '.join(df['title'].astype(str))

    # Tokenize the Greek titles and remove custom Greek stop words
    words = nltk.word_tokenize(titles)
    words = [word.lower() for word in words if word.isalnum() and word.lower() not in greek_stopwords]

    # Join the filtered words back into a single string
    filtered_titles = ' '.join(words)

    # Generate a word cloud
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate(filtered_titles)

    # Display or save the word cloud
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.title(f"Word Cloud for {sheet_name}")
    plt.tight_layout()

    # Save the word cloud image
    output_path = os.path.join(output_dir, f"{sheet_name}_wordcloud.png")
    plt.savefig(output_path)
    plt.close()

    print(f"Word cloud saved for sheet: {sheet_name}")

print("Word clouds generated for all sheets.")
```

Εικόνα 61: Κώδικας για την υλοποίηση των γραφημάτων word cloud ανά κατηγορία.

Παράρτημα-4



```python
# Import the necessary libraries
import pandas as pd
import matplotlib.pyplot as plt

# Load the Excel file into a pandas DataFrame
xls = pd.ExcelFile('total_duplicates_categorization.xlsx')
dfs = {sheet_name: xls.parse(sheet_name) for sheet_name in xls.sheet_names}

# Create a list to store daily counts
daily_counts_list = []

# Iterate through the date values (ignoring the time) for all sheets
for sheet_name, df in dfs.items():
    if 'pubdate' in df.columns:
        # Convert the 'pubdate' column to datetime
        if df['pubdate'].dtype != 'datetime64[ns]':
            df['pubdate'] = pd.to_datetime(df['pubdate'])

        # Iterate through the date values (ignoring the time) for the current sheet
        for date in df['pubdate'].dt.date.unique():
            # Filter data for the current date (ignoring the time) in the current sheet
            df_date = df[df['pubdate'].dt.date == date]

            # Calculate the count for the current sheet for the current date
            sheet_count = len(df_date)

            # If it's 'test_df', store it in 'test_df' column; otherwise, store it in 'Other_Sheets' column
            if sheet_name == 'test_df':
                daily_counts_list.append({'Date': date, 'test_df': sheet_count, 'Other_Sheets': 0})
            else:
                # Check if the date already exists in the list
                existing_entry = next((entry for entry in daily_counts_list if entry['Date'] == date), None)
                if existing_entry:
                    existing_entry['Other_Sheets'] += sheet_count
                else:
                    daily_counts_list.append({'Date': date, 'test_df': 0, 'Other_Sheets': sheet_count})

# Create a DataFrame from the list of dictionaries
daily_counts = pd.DataFrame(daily_counts_list)

# Sort the DataFrame by date
daily_counts = daily_counts.sort_values(by='Date')

# Plot the bar chart
plt.figure(figsize=(12, 6))
width = 0.4  # Width of each bar

# Calculate the positions for the bars
x = range(len(daily_counts))

# Plot 'test_df' bars on the left
bars1 = plt.bar(x, daily_counts['test_df'], color='#01154E', width=width, label='test_df')

# Plot 'Other Sheets' bars on the right, shifted by the width
bars2 = plt.bar([i + width for i in x], daily_counts['Other_Sheets'], color='#0395DF', width=width, label='Other Sheets')

# Set the x-axis labels to be the dates
plt.xticks([i + width / 2 for i in x], daily_counts['Date'], rotation=45)

plt.xlabel('Date')
plt.ylabel('Count')
plt.title('Bar Plot with non-filtered and filtered events for Each Day')
plt.legend()

# Display numbers on top of each bar
for bar1, bar2 in zip(bars1, bars2):
    height1 = bar1.get_height()
    height2 = bar2.get_height()
    plt.text(bar1.get_x() + bar1.get_width() / 2, height1, f'{int(height1)}', ha='center', va='bottom')
    plt.text(bar2.get_x() + bar2.get_width() / 2, height2, f'{int(height2)}', ha='center', va='bottom')

plt.tight_layout()
# Save the plot to a file if needed
plt.savefig(f'Number-of-non-filtered-and-filtered-events_barplot.png')
plt.show()
```

Εικόνα 62: Κώδικας για την υλοποίηση του διαγράμματος σχετικά με το πλήθος φιλτραρισμένων και μη φιλτραρισμένων ειδήσεων ανά ημέρα

12 Επιστημονική βιβλιογραφία

[A02] Abiteboul, S. (2002, August). Issues in monitoring web data. In *Database and Expert Systems Applications: 13th International Conference, DEXA 2002 Aix-en-Provence, France, September 2–6, 2002 Proceedings* (pp. 1-8). Berlin, Heidelberg: Springer Berlin Heidelberg.

[ACLED] "ACLED: Bringing clarity to crisis", Available at: https://acleddata.com/curated-data-files/

[Alc23] SQLAlchemy authors and contributors (2023). The Python SQL Toolkit and Object Relational Mapper. *SQLAlchemy*. Available at: https://www.sqlalchemy.org/

[AZHL22] Alhamadani, A., Zhang, X., He, J., & Lu, C. T. (2022). LANS: Large-scale Arabic News Summarization Corpus. *arXiv preprint arXiv:2210.13600*.

[B+21] Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19-42.

[B23b] Bayer Mike (2023). SQLAlchemy 2.0.20. *Python Software Foundation*. Available at: https://pypi.org/project/SQLAlchemy/

[B23c] B. Richard (2023). What Is MySQL and How Does It Work. Available at: https://www.hostinger.com/tutorials/what-is-mysql

[BSH21] Barua, A., Sharif, O., & Hoque, M. M. (2021). Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation. *Procedia Computer Science*, *193*, 112-121.

[C20] Cepalia A. (2020). A Beginner's Guide to pip; Scripts, Modules, Packages, and Libraries. *Real Python*. Available at: https://realpython.com/lessons/scripts-modules-packages-and-libraries/

[CFI20] CFI Team (2020). White-Collar Crime. Available at: https://corporatefinanceinstitute.com/resources/esg/white-collar-crime/

[CW15] Croicu, M., & Weidmann, N. B. (2015). Improving the selection of news reports for event coding using ensemble classification. *Research & Politics*, *2*(4), 2053168015615596.

[D22] divyasri7702 (2022). Python pytz. *GeeksforGeeks.* Available at: https://www.geeksforgeeks.org/python-pytz/#article-meta-div

[DS22] Devarajan, V., & Subramanian, R. (2022). Analyzing semantic similarity amongst textual documents to suggest near duplicates. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(3), 1703-1711.

[DZ11] Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications*, *28*(2), 37-40.

[F06] Fairon, C. (2006). Corporator: A tool for creating RSS-based specialized corpora. In *Proceedings of the 2nd International Workshop on Web as Corpus*.

[FA07] Feng, A., & Allan, J. (2007, November). Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 821-830).

[FA09] Feng, A., & Allan, J. (2009, November). Incident threading for news passages. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1307-1316).

[GA02] Galton, A., & Augusto, J. C. (2002). Two approaches to event definition. In *Database and Expert Systems Applications: 13th International Conference, DEXA 2002 Aix-en-Provence, France, September 2–6, 2002 Proceedings 13* (pp. 547-556). Springer Berlin Heidelberg.

[GN06] Garcia, I., & Ng, Y. K. (2006, November). Eliminating redundant and less-informative RSS news articles based on word similarity and a fuzzy equivalence relation. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)* (pp. 465-473). IEEE.

[H14] Holth Daniel (2014). wheel 0.24.0. *Python Software Foundation*. Available at: https://pypi.org/project/wheel/0.24.0/

[HKYU02] Hatano, K., Kinutani, H., Yoshikawa, M., & Uemura, S. (2002, August). Information retrieval system for XML documents. In *Database and Expert Systems Applications: 13th International Conference, DEXA 2002 Aix-en-Provence, France, September 2–6, 2002 Proceedings* (pp. 758-767). Berlin, Heidelberg: Springer Berlin Heidelberg.

[HSE22] HSEWatch Health Safety & Environment Encyclopdia (2022). Types Of Accidents (With Comprehensive Classification). Available at: https://hsewatch.com/types-of-accidents/

[Jus22] Violent Crimes Defined by Law (2022). Available at: https://www.justia.com/criminal/offenses/violent-crimes/

[K19] Knerl L. (2019). What are Computer Drivers?. *HP Development Company, L.P*. Available at: https://www.hp.com/us-en/shop/tech-takes/what-are-computer-drivers

[K21] Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3).

[KB16] Kaur, G., & Bajaj, K. (2016). News classification and its techniques: a review. *IOSR Journal of Computer Engineering*, *18*(1), 22-26.

[KJS20] Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping.

[KM23] Katari, R., & Myneni, M. B. (2020, March). A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-5). IEEE.

[L17] Loupasakis Andreas (2017). greek-stemmer 0.1.1. *Python Software Foundation*. Available at: https://pypi.org/project/greek-stemmer/

[L19] Labots Matthijs (2019). os-sys 2.1.4. *Python Software Foundation*. Available at: https://pypi.org/project/os-sys/

[LRS19] Lawyer Referral Service of Central Texas (2019). What are the Types of Property Crime? Available at: https://austinlrs.com/blog/what-are-the-types-of-property-crime/

[LS13] Leetaru, K., & Schrodt, P. A. (2013, April). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention* (Vol. 2, No. 4, pp. 1-49). Citeseer.

[M22] Misra, R. (2022). News category dataset. *arXiv preprint arXiv:2209.11429*.

[MMPD] “Mass Mobilization Protest Dataset”, Available at: https://massmobilization.github.io/about.html

[NHS23] UK National Health Service (2023). A to Z list of common illnesses and conditions, causes and treatments. *NHS inform.* Availabe at: https://www.nhsinform.scot/illnesses-and-conditions/a-to-z

[NltkT23] NLTK Team (2023). nltk 3.8.1. *Python Software Foundation*. Available at: https://pypi.org/project/nltk/

[Ora23a] ORACLE (2023). MySQL Connectors. Available at: https://www.mysql.com/products/connector/

[Ora23b] What is MySQL? (2023). Available at: https://www.oracle.com/mysql/what-is-mysql/

[Ora23c] MySQL Workbench. Enterprise Edition (2023). Available at: https://www.mysql.com/products/workbench/

[P21] Pechlivanis Konstantinos (2021). Greek-Stemmer. GitHub, Inc. Available at: https://github.com/kpech21/Greek-Stemmer

[PM15] Pilgrim M., McKee K. (2015). feedparser 5.2.0 documentation. *Python Software Foundation, Python Package Index*. Available at: https://pythonhosted.org/feedparser/introduction.html

[PRC21] Pape, R. A., Rivas, A. A., & Chinchilla, A. C. (2021). Introducing the new CPOST dataset on suicide attacks. *Journal of Peace Research*, *58*(4), 826-838.

[PSF23a] Python Software Foundation (2023). time — Time access and conversions. Available at: https://docs.python.org/3/library/time.html

[PSF23b] Python Software Foundation (2023). unicodedata — Unicode Database. Available at: https://docs.python.org/3/library/unicodedata.html

[R23] Richardson Leonard (2023). beautifulsoup4 4.12.2. *Python Software Foundation*. Available at: https://pypi.org/project/beautifulsoup4/

[RD23] Refsnes Data (2023). Python Requests Module. *W3Schools*. Available at: https://www.w3schools.com/python/module_requests.asp

[Ro23] Rose Erik (2023). more-itertools 10.1.0. *Python Software Foundation*. Available at: https://pypi.org/project/more-itertools/

[S23] Singh Taranjeet (2023). Natural Language Processing With spaCy in Python. *Real Python.* Available at: https://realpython.com/natural-language-processing-spacy-python/

[SH98] Schrodt, P. A., & Hall, B. (1998). Kansas Event Data System (KEDS). *Unpublished paper*.

[SM13] Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, *50*(4), 523-532.

[SM92] Sapir, D. G., & Misson, C. (1992). The development of a database on disasters. *Disasters*, *16*(1), 74-80.

[Spa23] SpaCy (2023). Install spaCy. Explosion. Available at: https://spacy.io/usage

[TPDT23] The Pandas Development Team (2023). pandas 2.0.3. *Python Software Foundation*. Available at: https://pypi.org/project/pandas/

[V23] Duong Vu (2023). Generating WordClouds in Python Tutorial. Available at: https://www.datacamp.com/tutorial/wordcloud-python

[VRJD17] Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017, August). Document level sentiment analysis from news articles. In *2017 international conference on computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-4). IEEE.

[W+13] Ward, M. D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., & Radford, B. (2013). Comparing GDELT and ICEWS event data. *Analysis*, *21*(1), 267-297.

[WKGS19] Wang, Q., Kanagal, B., Garg, V., & Sivakumar, D. (2019, November). Constructing a comprehensive events database from the web. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 229-238).

[XG23] X. Charlie, Gazoni Eric (2023). openpyxl 3.1.2. *Python Software Foundation*. Availabe at: https://pypi.org/project/openpyxl/

[A22] Αετοπούλου Β. Α., (2022). Σχεδιασμός και Υλοποίηση Ευφυούς Πληροφοριακού Συστήματος Ανάλυσης και Διαχείρισης Ειδησεογραφικών

Δεδομένων από Ελληνικά Μέσα Ηλεκτρονικής Ενημέρωσης. *MSc Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα,* Ελληνικό Ανοιχτό Πανεπιστήμιο, Πάτρα.

[IBE04] Ίδρυμα της Βουλής των Ελλήνων για τον Κοινοβουλευτισμό και τη Δημοκρατία (2004). Available at: https://foundation.parliament.gr/el

[M23] Ρόζμαρι Μάντζαρη: "Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: ανίχνευση αναδημοσιευμένων εκδόσεων ειδήσεων και ειδήσεων που περιγράφουν την εξέλιξη σημαντικών γεγονότων", Πτυχιακή Εργασία, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας, Οκτώβριος 2023.