ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: ανίχνευση αναδημοσιευμένων εκδόσεων ειδήσεων και ειδήσεων που περιγράφουν την εξέλιξη σημαντικών γεγονότων

Ρόζμαρι Μάντζαρη

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Θεόδωρος Τζουραμάνης

Επίκουρος Καθηγητής

Λαμία, Οκτώβριος 2023

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: ανίχνευση αναδημοσιευμένων εκδόσεων ειδήσεων και ειδήσεων που περιγράφουν την εξέλιξη σημαντικών γεγονότων

Ρόζμαρι Μάντζαρη

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Θεόδωρος Τζουραμάνης

Επίκουρος Καθηγητής

Λαμία, Οκτώβριος 2023

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: ανίχνευση αναδημοσιευμένων εκδόσεων ειδήσεων και ειδήσεων που περιγράφουν την εξέλιξη σημαντικών γεγονότων

Μάντζαρη Ρόζμαρι

Τριμελής Επιτροπή:

Θεόδωρος Τζουραμάνης, Επίκουρος Καθηγητής (επιβλέπων)

Σωτήριος Τασουλής, Επίκουρος Καθηγητής

Κωνσταντίνος Δελήμπασης, Καθηγητής

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή εργασία πραγματοποιήθηκε στο Πανεπιστήμιο Θεσσαλίας, στο Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική.

Η ολοκλήρωση της πτυχιακής αυτής εργασίας θα ήταν αδύνατη χωρίς την πολύτιμη υποστήριξη και καθοδήγηση του καθηγητή μου, Τζουραμάνη Θεόδωρου. Του εκφράζω τις ειλικρινείς μου ευχαριστίες για την βοήθεια και τη τεχνογνωσία που μου προσέφερε.

Θέλω να εκφράσω ένα μεγάλο ευχαριστώ στην καλή μου φίλη και συνερευνήτριά μου, Σπυριδούλα Σκορδίλη για την άριστη συνεργασία που είχαμε στα πλαίσια αυτής της εργασίας, ως μέρος ενός μεγαλύτερου έργου που υλοποιήσαμε μαζί, δηλαδή της ανάπτυξης της πρώτης (ημι)αυτοματοποιημένης ελληνικής βάσης δεδομένων παγκόσμιων ειδησεογραφικών γεγονότων. Ευχαριστώ για την συνεχή επικοινωνία και βοήθεια, που ποτέ δε δίστασε να μου δώσει. Η αφοσίωση και η υποστήριξη της ήταν ανεκτίμητη και είμαι ευγνώμων που μοιραστήκαμε αυτή την εμπειρία.

Θέλω, επίσης, να ευχαριστήσω βαθύτατα την οικογένεια μου, τους φίλους μου και τα υπόλοιπα σημαντικά άτομα στη ζωή μου, για τη συνεχή συμπαράσταση, βοήθεια και κατανόηση που έδειξαν κατά τη διάρκεια της εκπόνησης της παρούσας πτυχιακής εργασίας, αλλά και καθ’ όλη τη διάρκεια των προπτυχιακών μου σπουδών.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Περιεχόμενα

Μέρος Α: Θεωρητικό υπόβαθρο και σχετικές εργασίες 1

1 Εισαγωγή – Στόχος της πτυχιακής εργασίας 1

1.1 Εισαγωγή 1

1.2 Στόχος πτυχιακής 2

1.3 Η έννοια του γεγονότος 3

2 Πληροφοριακά συστήματα καταγραφής ειδησεογραφικών γεγονότων σε πραγματικό χρόνο 6

2.1 Εισαγωγή 6

2.2 Integrated Conflict Early Warning System (ICEWS) 6

2.3 Social Conflict Analysis Database (SCAD) 8

2.4 The UCDP Georeferenced Event Dataset (UCDP GED) 10

2.5 Global Terrorism Database (GTD) 11

2.6 The Geocoded Disasters (GDIS) dataset 13

2.7 Επίλογος 15

3 Μέθοδοι εντοπισμού επανάληψης δημοσίευσης ειδησεογραφικών άρθρων 16

3.1 Εισαγωγή 16

3.2 Τεχνικές ανίχνευσης διπλοτύπων και σχεδόν διπλοτύπων 18

3.3 Επίλογος 34

4 Μέθοδοι εντοπισμού ειδησεογραφικών άρθρων που αναφέρονται στην εξέλιξη στο χρόνο γνωστών γεγονότων 36

4.1 Εισαγωγή 36

4.2 Ανάλυση 37

4.3 Επίλογος 44

Μέρος Β: Το προτεινόμενο πληροφοριακό σύστημα καταγραφής ειδησεογραφικών γεγονότων στα Ελληνικά 45

5 Ανάλυση και σχεδίαση του νέου συστήματος 45

5.1 Εισαγωγή 45

5.2 Συνοπτική περιγραφή του πληροφοριακού συστήματος 45

5.3 Τρόπος λειτουργίας Πληροφοριακού Συστήματος 47

5.4 Ανάλυση modules, πακέτων, βιβλιοθηκών, drivers και frameworks του κώδικα σε γλώσσα προγραμματισμού Python 49

5.5 Επίλογος 54

6 Δημιουργία βάσης δεδομένων 55

6.1 Εισαγωγή 55

6.2 MySQL και Workbench 55

6.3 Διαδικασία δημιουργίας και σύνδεσης της βάσης με τον κώδικα 57

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

6.4 Επίλογος 62
7 Εντοπισμός πανομοιότυπων ή σχεδόν όμοιων ειδησεογραφικών άρθρων 63
7.1 Φιλτράρισμα πανομοιότυπων ή σχεδόν όμοιων ειδήσεων στο DataFrame βάσει τίτλου και περιγραφής 63
7.2 Φιλτράρισμα πανομοιότυπων ή σχεδόν όμοιων ειδήσεων στη βάση δεδομένων..... 75
8 Εντοπισμός ειδησεογραφικών άρθρων που περιγράφουν νεότερες εξελίξεις για γνωστά γεγονότα 83
8.1 Εισαγωγή 83
8.2 Εντοπισμός ειδησεογραφικών άρθρων εξέλιξης γεγονότος 84
8.3 Επίλογος 92
9 Πειραματική μελέτη 94
9.1 Εισαγωγή 94
9.2 Πλήθος γεγονότων ανά κατηγορία 94
9.3 Γράφημα Word Cloud διπλοτύπων γεγονότων ανά κατηγορία 97
9.4 Γράφημα Word Cloud πρωτότυπων ειδήσεων ανά κατηγορία 99
9.5 Πλήθος διπλότυπων και πρωτότυπων ειδήσεων ανά ημέρα 102
9.6 Μετρικές αξιοπιστίας του συστήματος για την εύρεση διπλοτύπων ειδησεογραφικών γεγονότων 105
9.7 Πλήθος follow up ανά κατηγορία 107
9.8 Σχεσιακός Γράφος γεγονότος follow-up είδησης 109
9.9 Μετρικές αξιοπιστίας του συστήματος για την εύρεση follow up ειδησεογραφικών γεγονότων 110
10 Συμπεράσματα και προτάσεις για μελλοντική εργασία 114
10.1 Συμπεράσματα 114
10.2 Προτάσεις για μελλοντική εργασία 114
Παραρτήματα: Πηγαίος Κώδικας για τη Δημιουργία Γραφημάτων 118
1 Γράφημα Word Cloud 118
2 Γράφημα Word Cloud ανά ημέρα 119
3 Γράφημα πλήθους ειδησεογραφικών γεγονότων ανά κατηγορία 120
Επιστημονική βιβλιογραφία 121

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Μέρος Α: Θεωρητικό υπόβαθρο και σχετικές εργασίες

1 Εισαγωγή – Στόχος της πτυχιακής εργασίας

1.1 Εισαγωγή

Η πρόοδος στον τομέα των Διαδικτυακών Τεχνολογιών τα τελευταία δέκα χρόνια έχει επιφέρει σημαντική αύξηση στον αριθμό των ειδησεογραφικών ιστοσελίδων [A02][CFGM02]. Αυτή η αύξηση έχει επηρεάσει τις προτιμήσεις των χρηστών, οι οποίοι πλέον τάσσονται υπέρ της αναζήτησης ειδήσεων στο Διαδίκτυο, το οποίο προσφέρει ένα ευρύ φάσμα πληροφοριών σε χαμηλό κόστος [MGB14].

Ο όγκος των διαθέσιμων δεδομένων στον Παγκόσμιο Ιστό είναι τεράστιος και μετράται σε Zettabytes, δηλαδή δισεκατομμύρια Gigabytes [KJS20]. Αυτό συμβαίνει διότι τα δεδομένα στον Ιστό δεν είναι στατικά αλλά αντιθέτως, παράγονται και τροποποιούνται με εξαιρετική ταχύτητα. Σύμφωνα με τη στατιστική έκθεση της Διεθνούς Εταιρείας Δεδομένων (International Data Corporation (IDC)), εκτιμάται ότι μέχρι το τέλος του 2024 θα φτάσει τα 149 Zettabytes, προκαλώντας προκλήσεις στη διατήρηση των ψηφιακών δεδομένων στο μέλλον [DS22].

Πιο συγκεκριμένα, ο Παγκόσμιος Ιστός περιλαμβάνει πάνω από 50 δισεκατομμύρια σελίδες στο Διαδίκτυο [MGB14]. Επιπλέον, το 90% των πληροφοριών που παράγονται σήμερα δημιουργούνται σε ψηφιακή μορφή, τάση η οποία πρόκειται να αυξηθεί ακόμα περισσότερο στο μέλλον. [MGB14][DS22].

Είναι πλέον γεγονός ότι οι χρήστες τείνουν να διαβάζουν ειδήσεις από ιστοσελίδες που συγκεντρώνουν έγγραφα από διάφορες πηγές και παρουσιάζουν τα δεδομένα αυτά κατηγοριοποιημένα. Κατά τη διαδικασία αυτή, η δημιουργία εν μέρει ή εντελώς πανομοιότυπων εγγράφων είναι αναπόφευκτη, καθώς οι ειδησεογραφικοί ιστότοποι αναδημοσιεύουν ειδήσεις που προέρχονται από άλλες ιστοσελίδες είτε κάνοντας μικρές αλλαγές στο κείμενο είτε διατηρώντας το ανέπαφο.

Αυτή η αυξημένη ποσότητα περιττών πληροφοριών, λόγω της αυξημένης παραγωγής ψηφιακών εγγράφων, δημιουργεί το πρόβλημα των διπλότυπων ή σχεδόν διπλότυπων ειδήσεων στο Διαδίκτυο, το οποίο αποτελεί ένα από τα πλέον κοινά προβλήματα. [X+11]. Όσον αφορά τον ορισμό της διπλότυπης πληροφορίας, υπάρχουν διάφορες προσεγγίσεις, καθώς ο ορισμός του τι συνιστά αντίγραφο είναι ασαφής. Κατά

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

μία προσέγγιση, ένα αντίγραφο ορίζεται ως ένα έγγραφο που περιέχει τους ακριβείς συντακτικούς όρους χωρίς διαφορές στη μορφοποίηση. Γενικά, εάν ένα έγγραφο περιέχει περίπου το ίδιο σημασιολογικό περιεχόμενο, θεωρείται διπλότυπο, ανεξαρτήτως της ακριβούς συντακτικής ταύτισης [CFGM02][HYK10].

Γίνεται επομένως κατανοητό ότι τα διπλότυπα δεδομένα, εξαιτίας της φύσης τους, μπορούν να επιφέρουν αρκετά σημαντικά μειονεκτήματα. Τα διπλότυπα έγγραφα δημιουργούν πλεονασμό και συχνά οδηγούν σε απώλεια αποδοτικότητας πληροφοριακών συστημάτων. Πολλές τεχνικές έχουν αναπτυχθεί για να εξάγουν, να ανακτούν, να φιλτράρουν και να κατηγοριοποιούν πληροφορίες, αυξάνοντας έτσι τη σημασία της ανίχνευσης των διπλότυπων εγγράφων [DS22].

Επομένως, είναι προφανές ότι η ανίχνευση των σχεδόν διπλότυπων δεδομένων (Near-Duplicate Detection (NDD)) αποτελεί ένα σημαντικό ερευνητικό πρόβλημα και έχει μεγάλη σημασία στη βελτίωση της αποτελεσματικότητας της αναζήτησης και της αποθήκευσης πληροφοριών [HYK10]. Η εξάλειψη των περιττών ειδήσεων μπορεί να βοηθήσει στην εξοικονόμηση χρόνου όσον αφορά τον εντοπισμό χρήσιμων πληροφοριών και την αποφυγή της επαναλαμβανόμενης διαδικασίας φιλτραρίσματος αναπαραγόμενων πληροφοριών. Με αυτόν τον τρόπο, ενισχύεται η ποικιλομορφία των δεδομένων που παρέχονται στους χρήστες [GN06].

Από τα παραπάνω, λοιπόν, ανάγεται το συμπέρασμα πως η δημιουργία μιας ελληνικής βάσης δεδομένων ειδησεογραφικών γεγονότων από όλο τον κόσμο θα χρησιμεύσει στη διευκόλυνση της διαχείρισης των ειδησεογραφικών δεδομένων [WKGS19][U09].

1.2 Στόχος πτυχιακής

Η παρούσα πτυχιακή εργασία αποσκοπεί στη δημιουργία μιας ελληνικής βάσης δεδομένων ειδησεογραφικών γεγονότων από όλο τον κόσμο, η οποία θα αποτελέσει ένα σημαντικό εργαλείο για την παρακολούθηση και την ανάλυση γεγονότων. Σε έναν κόσμο όπου η πληροφορία ρέει γρήγορα και αδιάκοπα, η δημιουργία μιας τέτοιας βάσης δεδομένων αποτελεί αναγκαιότητα για την κατανόηση των ειδησεογραφικών συμβάντων.

Στο πλαίσιο αυτού του έργου, η συνερευνήτριά μου Σπυριδούλα Σκορδίλη ανέλαβε στην εργασία της [Σ23] το σημαντικό καθήκον της συλλογής, αναγνώρισης, επεξεργασίας και κατηγοριοποίησης των ειδησεογραφικών γεγονότων, ενισχύοντας

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

έτσι την βάση δεδομένων μας με υψηλής ποιότητας δεδομένα. Από την άλλη πλευρά, η δική μου έρευνα εστιάζει στην ανίχνευση αναδημοσιευμένων εκδόσεων ειδησεογραφικών άρθρων και ειδησεογραφικών άρθρων που περιγράφουν την εξέλιξη στο χρόνο σημαντικών γεγονότων. Συγκεκριμένα, εφαρμόστηκαν πολλαπλοί έλεγχοι διπλοτύπων, τόσο κατά την διάρκεια εκτέλεσης του κώδικα συλλογής και ανάλυσης των γεγονότων, όσο και κατά την αποθήκευση των γεγονότων στη βάση δεδομένων. Για την παρακολούθηση της εξέλιξης των γεγονότων στο χρόνο, αξιοποιήθηκαν τεχνικές ανίχνευσης και προσδιορισμού επώνυμων οντοτήτων στα ειδησεογραφικά άρθρα.

Η ανάπτυξη των διαδικασιών αυτών, πραγματοποιήθηκε σε γλώσσα προγραμματισμού Python. Ακόμη, για την αποθήκευση και τη διαχείριση των ειδησεογραφικών δεδομένων, χρησιμοποιήθηκε σύστημα διαχείρισης βάσεων δεδομένων MySQL. Αυτό το σύστημα επιτρέπει την αποτελεσματική αποθήκευση, ανάκτηση και ανάλυση των ειδησεογραφικών δεδομένων, ενισχύοντας την εφαρμογή μας με έναν αξιόπιστο και αποδοτικό τρόπο διαχείρισης της πληροφορίας.

Η συνεργασία με τη Σ. Σκορδίλη εμπλούτισε το έργο μας με πολύτιμα στοιχεία και δίνει τη δυνατότητα για μια πιο εκτεταμένη ανάλυση της ειδησεογραφίας. Το αποτέλεσμα αυτής της συνεργασίας θα προσφέρει ένα ολοκληρωμένο (ημι)αυτοματοποιημένο εργαλείο που μπορεί να χρησιμοποιηθεί όχι μόνο από εμάς, αλλά και από την ευρύτερη κοινότητα για την παρακολούθηση, την ανάλυση και την κατανόηση των ειδησεογραφικών εξελίξεων στην Ελλάδα και στο εξωτερικό.

Συνολικά, το κοινό έργο στοχεύει στη δημιουργία ενός συνεκτικού και οργανωμένου περιβάλλοντος για την αποθήκευση και διαχείριση ελληνικών ειδησεογραφικών δεδομένων, με σκοπό τη βελτίωση της πρόσβασης των χρηστών σε σημαντικές ειδήσεις και την αποτροπή της επανάληψης των ίδιων πληροφοριών.

1.3 Η έννοια του γεγονότος

Όπως αναφέρει εκτενώς η συνεργάτιδά μου, δεν έχει καθοριστεί ένας συγκεκριμένος ορισμός για τη λέξη «γεγονός», όμως υπάρχουν διάφορες ερμηνείες με βάση τη βιβλιογραφία. Μερικές από αυτές είναι:

> «Κάτι που συμβαίνει σε μια συγκεκριμένη χρονική στιγμή ή σε μια συγκεκριμένη περίσταση, ανεξάρτητα από το αν είναι σπάνιο, συνηθισμένο ή τυχαίο.»

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

«Κάτι με ιδιαίτερη σημασία και σπουδαιότητα.»

«Κάτι συγκεκριμένο, πάντα αξιοσημείωτο.»

Ένα ειδησεογραφικό γεγονός είναι μια συγκεκριμένη γεγονοτική εξέλιξη ή γεγονός που περιγράφεται και αναφέρεται στα μέσα ενημέρωσης, όπως εφημερίδες, ιστοσελίδες, κ.λπ. Τα ειδησεογραφικά γεγονότα αποτελούν συχνά σημαντικά, ενδιαφέροντα συμβάντα που επηρεάζουν την κοινωνία, την πολιτική, την οικονομία, την πολιτισμική ζωή και άλλους τομείς. Τα ειδησεογραφικά γεγονότα παρέχουν πληροφορίες στο κοινό και μπορούν να αποτελέσουν το αντικείμενο δημοσιογραφικής έρευνας και ανάλυσης.

Κάθε ειδησεογραφικό γεγονός έχει ένα μοναδικό αναγνωριστικό, ένα κομμάτι κειμένου και τον χρόνο πηγής, που σηματοδοτεί πότε δημοσιεύτηκε. Επιπλέον, περιέχει ένα πεδίο κειμένου, το οποίο είναι η ένωση πολλών χαρακτήρων. Ένα απόσπασμα είναι ο συνδυασμός διαδοχικών χαρακτήρων που αντιπροσωπεύει μια πλήρη περιγραφή κάποιου γεγονότος ειδήσεων. Μια ιστορία μπορεί να περιέχει ένα ή περισσότερα αποσπάσματα. Ένα γεγονός περιγράφεται από τα εξής [FA07] [FA09]:

- Κύριοι χαρακτήρες (ΠΟΙΟΣ): Οι πιο σημαντικές ονομαστικές οντότητες σε ένα απόσπασμα που δείχνουν ποιος ή τι εμπλέκεται στο γεγονός που περιγράφεται από το απόσπασμα.
- Χρονικές σημάνσεις (ΠΟΤΕ): Λαμβάνουμε υπόψη δύο τύπους χρονικών χαρακτηριστικών. Ο ένας είναι ο χρόνος δημοσίευσης της είδησης, ο οποίος είναι ίδιος για όλα τα αποσπάσματα μιας ιστορίας. Το άλλο είναι το απόλυτο ή σχετικό χρονικό σημείο ή εύρος που αναφέρεται στο απόσπασμα όταν συνέβη το αντίστοιχο γεγονός.
- Τοποθεσία (ΠΟΥ): Η γεωγραφική θέση όπου συνέβη το γεγονός που περιγράφεται στο απόσπασμα. Είναι πολύ συνηθισμένο ότι οι πληροφορίες τοποθεσίας δεν είναι διαθέσιμες στο απόσπασμα και απαιτείται συμπέρασμα από το περιβάλλον υπό αυτήν την προϋπόθεση.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

- Δράση (ΤΙ): Το βασικό ρήμα που περιγράφει το πραγματικό γεγονός στο απόσπασμα.

Στις 10 Δεκεμβρίου 1963, απονεμήθηκε το βραβείο Νόμπελ Λογοτεχνίας, στον Γεώργιο Σεφέρη, στην τελετή απονομής που πραγματοποιήθηκε στη Στοκχόλμη. Ο Γ. Σεφέρης υπήρξε ο πρώτος Έλληνας που τιμήθηκε με βραβείο Νόμπελ.

ΠΟΙΟΣ ΠΟΤΕ ΠΟΥ ΤΙ

Εικόνα 1-1: Παράδειγμα γεγονότος με τις ιδιότητες που το περιγράφουν

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2 Πληροφοριακά συστήματα καταγραφής ειδησεογραφικών γεγονότων σε πραγματικό χρόνο

2.1 Εισαγωγή

Οι βάσεις δεδομένων διαδραματίζουν κρίσιμο ρόλο στην ανάλυση και τη διαχείριση γεγονότων, είτε αφορούν περιστατικά όπως αυτοκτονίες είτε πολιτικά σημαντικά συμβάντα. Με τη συλλογή, την οργάνωση και την ερμηνεία δεδομένων, οι βάσεις δεδομένων παρέχουν ένα δομημένο πλαίσιο για την κατανόηση και την αντιμετώπιση των δυσκολιών που σχετίζονται με τέτοια συμβάντα.

Η κεντρική διαχείριση δεδομένων είναι βασικό πλεονέκτημα των βάσεων δεδομένων. Μπορούν να αποθηκεύονται και να προσπελάσονται με ασφάλεια πληροφορίες που σχετίζονται με συμβάντα. Οι βάσεις δεδομένων απλοποιούν την αποτελεσματική ανάκτηση και παρέχουν μια ολοκληρωμένη εικόνα του συμβάντος, επειδή η πληροφορία είναι συγκεντρωμένη. Μετά από ένα συμβάν, η έγκαιρη πρόσβαση σε σχετικές πληροφορίες είναι κρίσιμη. Οι βάσεις δεδομένων διευκολύνουν τη γρήγορη και στοχευμένη ανάκτηση, βοηθώντας τους ερευνητές και τις αρχές να αποκτήσουν γνώσεις και να λάβουν τις απαραίτητες ενέργειες.

Με τη χρήση εργαλείων ανάλυσης δεδομένων , είναι δυνατό να εντοπιστούν μοτίβα, τάσεις και συσχετίσεις, προσφέροντας βαθύτερη κατανόηση των κρυφών παραγόντων και των πιθανών αιτίων που σχετίζονται με γεγονότα. Αυτή η ανάλυση συμβάλλει στην ενημέρωση των προληπτικών μέτρων, την ανάπτυξη πολιτικής και τις παρεμβάσεις. Επιπλέον, οι βάσεις δεδομένων υποστηρίζουν τη συνεργασία και την ανταλλαγή πληροφοριών μεταξύ των ενδιαφερομένων που συμμετέχουν στην ανάλυση και διαχείριση των γεγονότων. Πέραν αυτού, οι βάσεις δεδομένων διατηρούν ακριβή ιστορικά αρχεία γεγονότων, τα οποία αποτελούν πολύτιμη πηγή για μελλοντική αναφορά, έρευνα και ανάπτυξη πολιτικής.

Στις επόμενες ενότητες, θα εξερευνήσουμε διάφορους τύπους βάσεων δεδομένων και τον τρόπο χρήσης τους στην ανάλυση και διαχείριση των γεγονότων. Θα αναδείξουμε τα χαρακτηριστικά ορισμένων καλά γνωστών βάσεων δεδομένων, τονίζοντας τη σημαντικότητά τους στον επιστημονικό και ερευνητικό τομέα.

2.2 Integrated Conflict Early Warning System (ICEWS)

Το Integrated Conflict Early Warning System (ICEWS) είναι μια βάση δεδομένων που συλλέγει και αναλύει δεδομένα για πολιτικά γεγονότα και συγκρούσεις

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

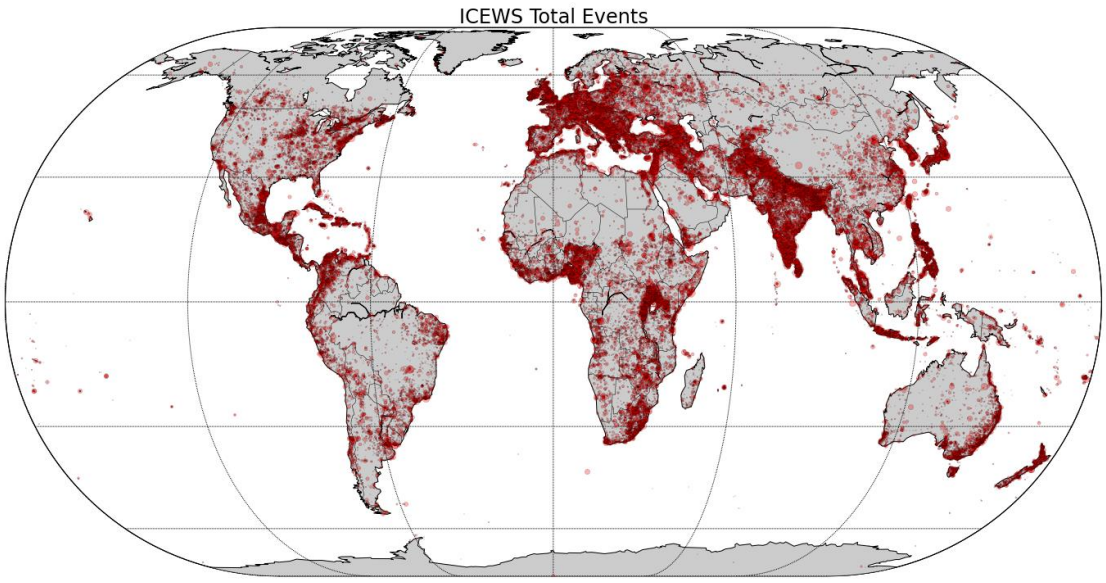
παγκοσμίως. Τα ICEWS Weekly Event Data είναι ένα υποσύνολο αυτής της βάσης δεδομένων που παρέχει πληροφορίες για παγκόσμια συμβάντα που συνέβησαν σε μια δεδομένη εβδομάδα. Είναι ένα δημοσίως διαθέσιμο σύνολο δεδομένων που περιλαμβάνει πληροφορίες για πολιτικά γεγονότα όπως διαμαρτυρίες, βία και διπλωματικές δραστηριότητες. Επιπλέον, παρέχει μια ποικιλία πληροφοριών για κάθε συμβάν, συμπεριλαμβανομένης της ημερομηνίας, της τοποθεσίας, των εμπλεκόμενων, του είδους του συμβάντος και των αποτελεσμάτων.

Ορισμένα βασικά χαρακτηριστικά των Weekly Event Data του ICEWS περιλαμβάνουν:

1. *Γεωγραφική κάλυψη*: Το σύνολο δεδομένων καλύπτει παγκόσμια γεγονότα και παρέχει μια ευρεία γεωγραφική αναπαράσταση, με περιοχές όπως η Αφρική, η Ασία, η Ευρώπη, η Λατινική Αμερική και η Μέση Ανατολή.
2. *Δεδομένα σε επίπεδο συμβάντων*: Το σύνολο δεδομένων παρέχει λεπτομερείς πληροφορίες για μεμονωμένα γεγονότα, επιτρέποντας τη σε βάθος ανάλυση της δυναμικής και των προτύπων των πολιτικών γεγονότων.
3. *Έγκαιρες ενημερώσεις*: Το σύνολο δεδομένων ενημερώνεται εβδομαδιαία, αντικατοπτρίζοντας τις πιο πρόσφατες εξελίξεις και διασφαλίζοντας τη συνάφειά του με τη συνεχή έρευνα και ανάλυση.
4. *Πρόσβαση και εργαλεία ανάλυσης*: Το σύνολο δεδομένων είναι ελεύθερα διαθέσιμο στο κοινό και οι χρήστες μπορούν να κατεβάσουν τα δεδομένα για τους δικούς τους ερευνητικούς σκοπούς. Το ICEWS παρέχει επίσης διαδικτυακά εργαλεία και επιλογές οπτικοποίησης για να διευκολύνει την εξερεύνηση και ανάλυση δεδομένων.

Τα ICEWS Weekly Event Data έχουν χρησιμοποιηθεί ευρέως από ερευνητές, υπεύθυνους χάραξης πολιτικής και οργανισμούς που ειδικεύονται στην ανάλυση συγκρούσεων και τις προσπάθειες εδραίωσης της ειρήνης. Παρέχει πολύτιμες γνώσεις για τα πολιτικά γεγονότα και τη δυναμική των συγκρούσεων, διευκολύνοντας την καλύτερη κατανόηση των παγκόσμιων γεγονότων και των επιπτώσεών τους στους πληγέντες πληθυσμούς. [W+13]

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Εικόνα 2-1: Παράδειγμα ανάλυσης ICEWS: Οι κόκκινες κουκίδες αντιπροσωπεύουν κάθε πολιτικό γεγονός που έχει καταγραφεί στη βάση δεδομένων ICEWS έως το 2015.

2.3 Social Conflict Analysis Database (SCAD)

Η Social Conflict Analysis Database (SCAD) είναι ένα ευρέως αναγνωρισμένο σύνολο δεδομένων που εστιάζει στην παρακολούθηση και ανάλυση γεγονότων κοινωνικών και πολιτικών συγκρούσεων παγκοσμίως. Το SCAD παρέχει λεπτομερείς πληροφορίες για διάφορες μορφές κοινωνικής σύγκρουσης, όπως διαμαρτυρίες, απεργίες και άλλους τύπους αμφιλεγόμενων γεγονότων.

Περιλαμβάνει δεδομένα για γεγονότα συγκρούσεων από διαφορετικές περιοχές και χρονικές περιόδους, επιτρέποντας έτσι, τις συγκριτικές αναλύσεις και τον εντοπισμό τάσεων και προτύπων. Παρέχει πληροφορίες για την τοποθεσία, την ημερομηνία, τις εμπλεκόμενες οντότητες, τις αιτίες, τις τακτικές και τα αποτελέσματα κάθε καταγεγραμμένου συμβάντος. Το SCAD στοχεύει στην καταγραφή ενός ευρέος φάσματος κοινωνικών και πολιτικών συγκρούσεων, επιτρέποντας στους ερευνητές να εξετάσουν τη δυναμική και τις συνέπειες αυτών των γεγονότων.

Ορισμένα από τα βασικά χαρακτηριστικά του SCAD αποτελούν:

1. *Πεδίο εφαρμογής και κάλυψη*: Το SCAD καλύπτει ένα ευρύ φάσμα κοινωνικών και πολιτικών γεγονότων συγκρούσεων από όλο τον κόσμο. Περιλαμβάνει τόσο εσωτερικές όσο και διεθνείς συγκρούσεις, που εκτείνονται σε διάφορες περιοχές και χρονικές περιόδους.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2. *Ολοκληρωμένες μεταβλητές*: Το SCAD καταγράφει μια ποικιλία μεταβλητών που σχετίζονται με συμβάντα σύγκρουσης. Αυτό μπορεί να περιλαμβάνει πληροφορίες για το είδος της σύγκρουσης, το επίπεδο βίας, τον αριθμό των συμμετεχόντων, τον στόχο της σύγκρουσης και την παρουσία εξωτερικών παρεμβάσεων.
3. *Διαχρονική προοπτική*: Το SCAD επιτρέπει τη διαχρονική ανάλυση καλύπτοντας ένα σημαντικό χρονικό διάστημα, δίνοντας έτσι τη δυνατότητα στους ερευνητές να εξετάσουν τις τάσεις, τις αλλαγές και τα πρότυπα στις κοινωνικές συγκρούσεις με την πάροδο του χρόνου.
4. *Πολλαπλοί παράγοντες*: Το SCAD λαμβάνει υπόψη τη συμμετοχή διαφόρων παραγόντων σε κοινωνικές συγκρούσεις. Παρακολουθεί όχι μόνο τις ενέργειες των διαδηλωτών αλλά και τις απαντήσεις των σωμάτων ασφαλείας, των κυβερνητικών αρχών και άλλων σχετικών παραγόντων.
5. *Αξιοπιστία και εγκυρότητα*: Το SCAD αναπτύσσεται χρησιμοποιώντας αυστηρές μεθοδολογίες και αντλεί από αξιόπιστες πηγές, όπως ρεπορτάζ ειδήσεων, ακαδημαϊκές μελέτες και άλλους αξιόπιστους παρόχους πληροφοριών.
6. *Ανοικτή πρόσβαση*: Το SCAD διατίθεται στο κοινό ή στην ακαδημαϊκή κοινότητα, επιτρέποντας στους ερευνητές, τους υπεύθυνους χάραξης πολιτικής και τους αναλυτές να έχουν πρόσβαση και να χρησιμοποιούν το σύνολο δεδομένων για τις δικές τους μελέτες και αναλύσεις.

Οι ερευνητές και οι αναλυτές χρησιμοποιούν το SCAD για να διερευνήσουν ένα ευρύ φάσμα θεμάτων που σχετίζονται με τις κοινωνικές συγκρούσεις, όπως των παραγόντων που οδηγούν στην κινητοποίηση, των στρατηγικών που χρησιμοποιούνται από διαφορετικούς φορείς και των αποτελεσμάτων και των συνεπειών αυτών των συγκρούσεων. Η βάση δεδομένων παρέχει μια πολύτιμη πηγή για τη μελέτη της δυναμικής των κοινωνικών συγκρούσεων και των επιπτώσεών της στις κοινωνίες.[S+12] [SCAD]

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Εικόνα 2-2: Παράδειγμα ανάλυσης SCAD: Γεωαναφερόμενες συλλήψεις για λαθρεμπόριο μεταναστών στο Sullivan City του Τέξας. Πηγή: Τμήμα Δημόσιας Ασφάλειας του Τέξας.

2.4 The Uppsala Conflict Data Program Georeferenced Event Dataset (UCDP GED)

Το UCDP Georeferenced Event Dataset (UCDP GED) είναι ένα σύνολο δεδομένων που παρέχει λεπτομερείς πληροφορίες για ένοπλες συγκρούσεις, πολιτική βία και άλλα σχετικά γεγονότα παγκοσμίως. Το UCDP GED εστιάζει στο γεωγραφικό εντοπισμό και την παροχή δεδομένων σε επίπεδο συμβάντων για διάφορους τύπους περιστατικών που σχετίζονται με συγκρούσεις, όπως μάχες, βομβαρδισμοί, ταραχές, διαμαρτυρίες και άλλες μορφές πολιτικής βίας. Καλύπτει τόσο διεθνείς όσο και εσωτερικές συγκρούσεις, που εκτείνονται από το 1989 έως σήμερα.

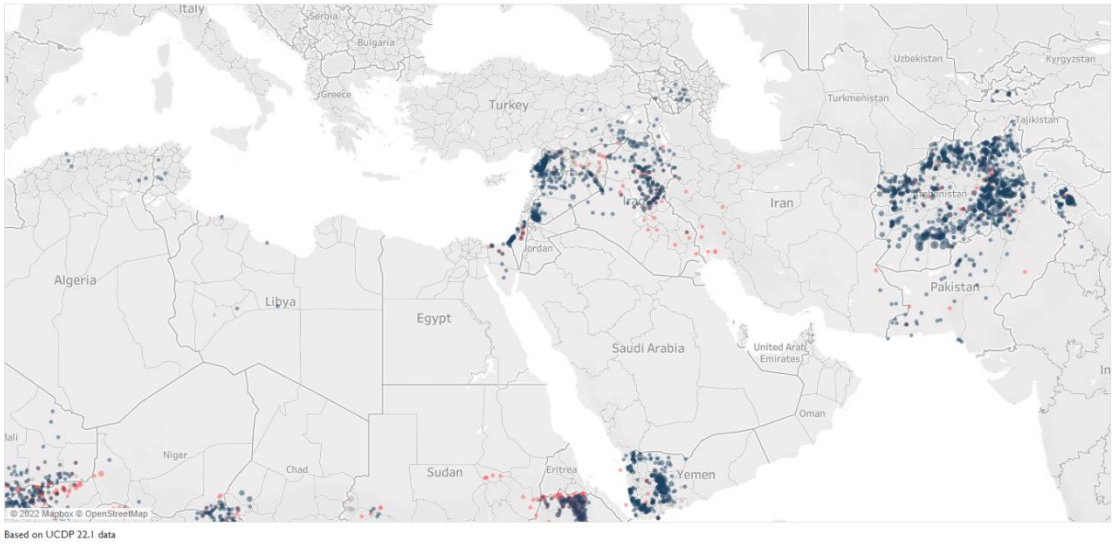
Τα βασικά χαρακτηριστικά του UCDP GED περιλαμβάνουν:

1. *Γεωγραφικές πληροφορίες:* Το σύνολο δεδομένων περιλαμβάνει ακριβείς συντεταγμένες με γεωαναφορά για κάθε καταγεγραμμένο συμβάν. Αυτό επιτρέπει χωρική ανάλυση, χαρτογράφηση και οπτικοποίηση γεγονότων σύγκρουσης.
2. *Λεπτομέρειες σε επίπεδο συμβάντος:* Το UCDP GED παρέχει πληροφορίες σχετικά με τα συγκεκριμένα χαρακτηριστικά κάθε συμβάντος: της ημερομηνίας, της τοποθεσίας, των εμπλεκόμενων παραγόντων, των τύπων βίας που ασκήθηκαν και των θυμάτων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

3. *Διαχρονική προοπτική:* Το σύνολο δεδομένων καλύπτει μια σημαντική χρονική περίοδο, επιτρέποντας στους ερευνητές και τους αναλυτές να εξετάσουν τις τάσεις, τα πρότυπα και τις αλλαγές στη δυναμική των συγκρούσεων με την πάροδο του χρόνου.
4. *Τακτικές ενημερώσεις:* Το UCDP GED ενημερώνεται τακτικά για να ενσωματώνει νέα συμβάντα και να περιλαμβάνει αναθεωρήσεις ή πρόσθετες πληροφορίες για υπάρχουσες καταχωρίσεις. Αυτό διασφαλίζει ότι το σύνολο δεδομένων παραμένει ενημερωμένο και πολύτιμο για συνεχή έρευνα.

Το UCDP GED χρησιμεύει ως πολύτιμος πόρος για ερευνητές, υπεύθυνους χάραξης πολιτικής και οργανισμούς που ενδιαφέρονται να κατανοήσουν και να αναλύσουν τις ένοπλες συγκρούσεις και την πολιτική βία. Επιτρέπει την εις βάθος ανάλυση της δυναμικής των συγκρούσεων, της χωρικής κατανομής της βίας και του αντίκτυπου των συγκρούσεων στους πληθυσμούς. [SM13]



Εικόνα 2-3: Παράδειγμα ανάλυσης UCDP: Θανατηφόρα γεγονότα του 2021 κατηγοριοποιημένα ανά τύπο βίας στην Μέση Ανατολή. Μπλε: Κρατική βία, Γκρι: Όχι κρατική βία, Πορτοκαλί: Μονόπλευρη βία. Το μέγεθος της κουκίδας αυξάνεται ανά απώλεια.

2.5 Global Terrorism Database (GTD)

Η Global Terrorism Database (GTD) αποτελεί μία βάση δεδομένων που παρέχει λεπτομερείς πληροφορίες για τρομοκρατικές επιθέσεις παγκοσμίως. Το GTD περιλαμβάνει δεδομένα τόσο για εγχώρια όσο και για διεθνή τρομοκρατικά περιστατικά από το 1970 έως σήμερα. Καλύπτει ένα ευρύ φάσμα τρομοκρατικών δραστηριοτήτων, όπως των βομβαρδισμών, των δολοφονιών, των αεροπειρατειών, των απαγωγών και άλλων μορφών επιθέσεων. Η βάση δεδομένων συλλέγει πληροφορίες

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

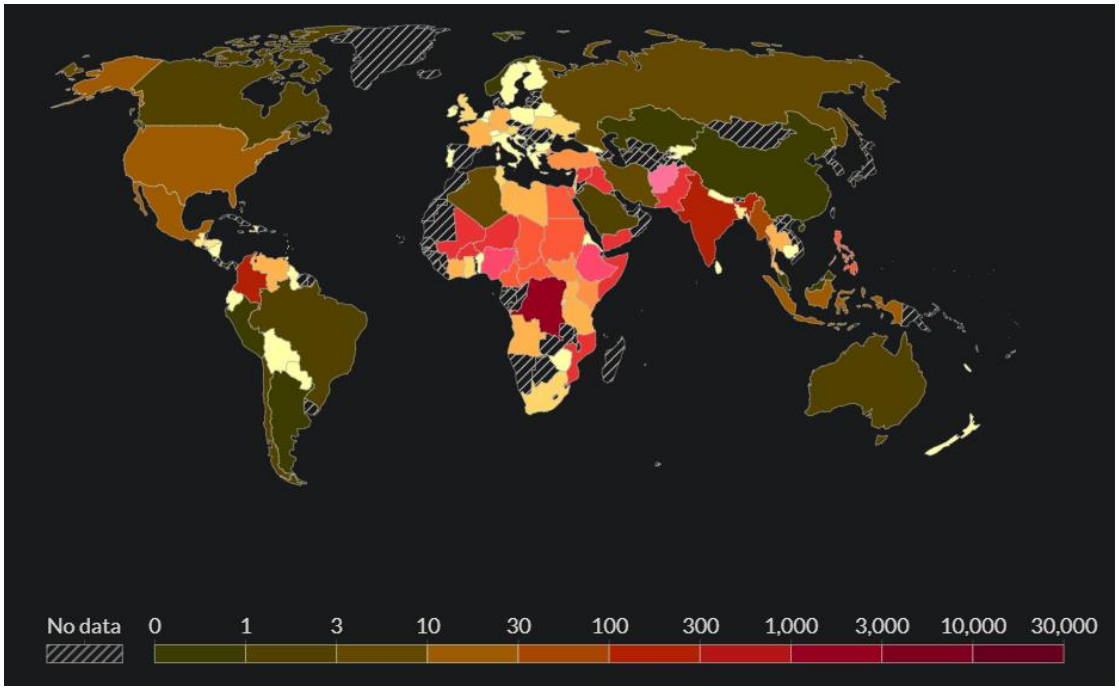
από διάφορες πηγές, όπως αναφορές μέσων ενημέρωσης, κρατικά αρχεία, ακαδημαϊκές έρευνες και συνεισφορές από μελετητές και ειδικούς στον τομέα.

Μερικά βασικά χαρακτηριστικά του GTD περιλαμβάνουν:

1. *Εκτεταμένη κάλυψη*: Η βάση δεδομένων παρέχει λεπτομερείς πληροφορίες για διάφορες πτυχές τρομοκρατικών συμβάντων, όπως η ημερομηνία, η τοποθεσία, τα όπλα που χρησιμοποιήθηκαν, οι τύποι στόχων, τα θύματα και οι δράστες.
2. *Εύρεση Μοτίβων*: Οι ερευνητές και οι αναλυτές μπορούν να χρησιμοποιήσουν το GTD για να εξετάσουν τις τάσεις, τα πρότυπα και τα χαρακτηριστικά της τρομοκρατίας παγκοσμίως. Επιτρέπει το φιλτράρισμα και την αναζήτηση βάσει διαφορετικών κριτηρίων για την εξερεύνηση των δεδομένων.
3. *Συνεχείς ενημερώσεις*: Το GTD ενημερώνεται τακτικά για να περιλαμβάνει νέα περιστατικά και να ενσωματώνει πρόσθετες πληροφορίες ή διορθώσεις για υπάρχουσες καταχωρίσεις. Αυτό διασφαλίζει ότι η βάση δεδομένων παραμένει ενημερωμένη και σχετική για συνεχή έρευνα και ανάλυση.
4. *Προσβασιμότητα*: Το GTD είναι ελεύθερα προσβάσιμο στο κοινό και παρέχει στους χρήστες τη δυνατότητα να κατεβάσουν τα δεδομένα και να τα χρησιμοποιήσουν για τους δικούς τους ερευνητικούς σκοπούς.

Ερευνητές, υπεύθυνοι χάραξης πολιτικής και αναλυτές χρησιμοποιούν το GTD για να αποκτήσουν γνώσεις σχετικά με τη δυναμική της τρομοκρατίας, να κατανοήσουν τις τάσεις και τα πρότυπα και να ενημερώσουν τις αντιτρομοκρατικές στρατηγικές και πολιτικές. [LD07]

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Εικόνα 2-4: Παράδειγμα ανάλυσης GTD: Θάνατοι από τρομοκρατία, 2020. Επιβεβαιωμένοι θάνατοι, συμπεριλαμβανομένων όλων των θυμάτων και των δραστών που έχασαν τη ζωή τους ως αποτέλεσμα του συμβάντος.

2.6 The Geocoded Disasters (GDIS) dataset

Το Geocoded Disasters (GDIS) είναι ένα σύνολο δεδομένων που παρέχει πληροφορίες για διάφορους τύπους καταστροφών που συμβαίνουν παγκοσμίως. Περιλαμβάνει ακριβείς γεωγραφικές συντεταγμένες που σχετίζονται με κάθε συμβάν καταστροφής, επιτρέποντας χωρική ανάλυση και χαρτογράφηση.

Η βάση δεδομένων GDIS καταγράφει ένα ευρύ φάσμα καταστροφών, από φυσικές καταστροφές όπως σεισμοί, πλημμύρες, τυφώνες, πυρκαγιές και ξηρασίες, εώς και ανθρωπογενείς καταστροφές όπως βιομηχανικά ατυχήματα ή τρομοκρατικές επιθέσεις. Στοχεύει στην παροχή ολοκληρωμένων και αξιόπιστων πληροφοριών σχετικά με την εμφάνιση, την τοποθεσία, τη σοβαρότητα και τις επιπτώσεις αυτών των καταστροφών.

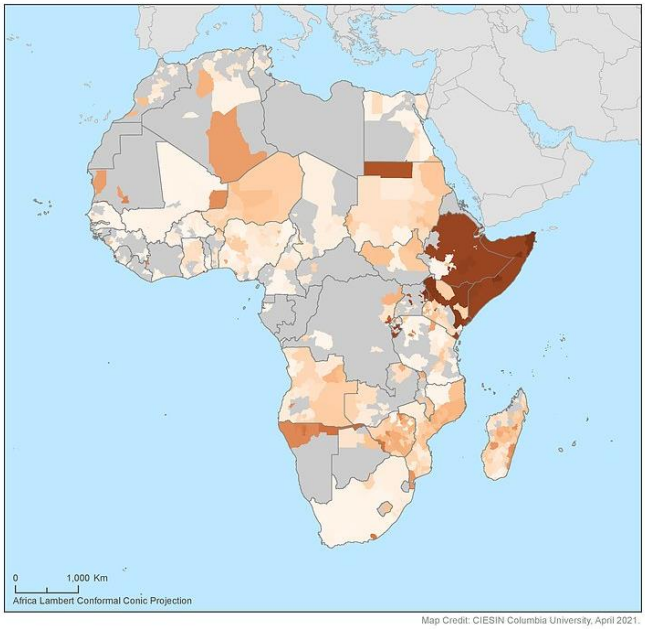
Το GDIS παρέχει πληροφορίες σε επίπεδο μεμονωμένων γεγονότων καταστροφής, προσφέροντας λεπτομέρειες σχετικά με την ημερομηνία, την ώρα, την τοποθεσία (γεωγραφικό πλάτος και μήκος), τον τύπο της καταστροφής και τα χαρακτηριστικά του συμβάντος.

Μερικά από τα βασικά χαρακτηριστικά του GDIS αποτελούν:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

1. *Χωρική ανάλυση*: Η γεωκωδικοποιημένη (geocoded) φύση της βάσης δεδομένων επιτρέπει τη χωρική ανάλυση και χαρτογράφηση γεγονότων καταστροφών. Οι ερευνητές μπορούν να οπτικοποιήσουν την κατανομή και την ομαδοποίηση των καταστροφών, να εντοπίσουν περιοχές υψηλού κινδύνου και να αναλύσουν χωρικά μοτίβα και σχέσεις.
2. *Ολοκληρωμένη κάλυψη*: Η βάση δεδομένων στοχεύει να καλύψει ένα ευρύ φάσμα καταστροφών από διάφορες περιοχές και χρονικές περιόδους, παρέχοντας μια συνολική εικόνα για τα περιστατικά καταστροφών.
3. *Εκτίμηση επιπτώσεων*: Το GDIS μπορεί να περιλαμβάνει πληροφορίες σχετικά με τη σοβαρότητα και τις επιπτώσεις των καταστροφών, όπως ο αριθμός των θυμάτων, ο πληγέντος πληθυσμός και η οικονομική ζημιά.
4. *Πηγές δεδομένων*: Η βάση δεδομένων GDIS αντλεί πληροφορίες από πολλές πηγές, όπως κυβερνητικές εκθέσεις, διεθνείς οργανισμούς, ακαδημαϊκές μελέτες, ειδησεογραφικά άρθρα και άλλες αξιόπιστες πηγές.

Το GDIS χρησιμοποιείται από ερευνητές, υπεύθυνους χάραξης πολιτικής, ανθρωπιστικές οργανώσεις και άλλους που ενδιαφέρονται να μελετήσουν τις καταστροφές και να αναπτύξουν στρατηγικές για τη διαχείριση και την αντιμετώπιση καταστροφών [RB21].



Εικόνα 2-5: Παράδειγμα ανάλυσης GDIS: Ποσοστό θνησιμότητας στην Αφρική από το 2009 έως το 2018 από φυσικές καταστροφές. Όσο πιο σκούρο είναι το χρώμα, τόσο μεγαλύτερη θνησιμότητα υπάρχει.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2.7 Επίλογος

Καθώς ολοκληρώνουμε την εξερεύνηση της σημασίας των βάσεων δεδομένων σε αυτόν τον τομέα, είναι προφανές ότι παίζουν ζωτικό ρόλο στην κατανόηση και την πρόληψη των επιπτώσεων των γεγονότων. Οι βάσεις δεδομένων έχουν δώσει τη δυνατότητα στους αναλυτές να εμβαθύνουν εκτενώς στα δεδομένα διάφορων γεγονότων, αποκαλύπτοντας μοτίβα, τάσεις και συνδέσεις που συμβάλλουν σε μια ολοκληρωμένη κατανόηση των γεγονότων. Η ικανότητα ανάλυσης και συσχέτισης δεδομένων έχει διευκολύνει τον εντοπισμό των υποκείμενων παραγόντων, των πιθανών αιτιών και των συνεπειών. Αυτή η γνώση έχει ανοίξει το δρόμο για την ανάπτυξη προληπτικών στρατηγικών, παρεμβάσεων πολιτικής και βελτιώσεων στις πρακτικές διαχείρισης εκδηλώσεων.

Εκτός από τις αναλυτικές τους ικανότητες, οι βάσεις δεδομένων παρέχουν ελεγχόμενη πρόσβαση και ασφαλείς μηχανισμούς ανταλλαγής δεδομένων. Επιπλέον, τα ιστορικά αρχεία που διατηρούνται σε βάσεις δεδομένων χρησιμεύουν ως πολύτιμη πηγή για μελλοντική αναφορά και έρευνα. Οι τεχνολογικές εξελίξεις, όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση, υπόσχονται περαιτέρω ενίσχυση της ισχύος και της αποτελεσματικότητας των βάσεων δεδομένων στην κατανόηση γεγονότων, την πρόβλεψη πιθανών κινδύνων και την καθοδήγηση της προληπτικής λήψης αποφάσεων.

Συμπερασματικά, η σημασία των βάσεων δεδομένων στην ανάλυση και διαχείριση συμβάντων δεν μπορεί να υπερεκτιμηθεί. Ο ρόλος τους στη συγκέντρωση δεδομένων, την αποτελεσματική ανάκτηση, την υποστήριξη της ανάλυσης και της συνεργασίας και τη διατήρηση ιστορικών αρχείων έχει αλλάξει τον τρόπο με τον οποίο προσεγγίζουμε τα γεγονότα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

3 Μέθοδοι εντοπισμού επανάληψης δημοσίευσης ειδησεογραφικών άρθρων

3.1 Εισαγωγή

Σε μια εποχή που χαρακτηρίζεται από υπερφόρτωση πληροφοριών και γρήγορο πολλαπλασιασμό των ψηφιακών ειδήσεων, ο εντοπισμός και ο μετριασμός σχεδόν διπλότυπων ειδησεογραφικών άρθρων έχουν αναδειχθεί ως επιτακτικές προκλήσεις. Ως συγκεκριμένο παράδειγμα, ένα σημαντικό ποσοστό των ιστοσελίδων βρέθηκε να είναι σχεδόν διπλότυπες από διάφορες μελέτες. Αυτές, δείχνουν ότι περίπου το 1,7% έως το 7% των ιστοσελίδων που επισκέπτονται οι ανιχνευτές είναι σχεδόν διπλότυπες σελίδες [X+11].

Τα σχεδόν διπλότυπα άρθρα αναφέρονται σε κομμάτια περιεχομένου που έχουν σημαντικές ομοιότητες όσον αφορά το περιεχόμενο κειμένου, τη δομή ή το συνολικό τους μήνυμα, ενώ παρουσιάζουν μικρές παραλλαγές, όπως τυπογραφικά λάθη, έγγραφα διαφορετικής έκδοσης, πιστής αντιγραφής ή λογοκλοπής κ.λπ. [X+11].

Η λογοκλοπή (plagiarism) στα ειδησεογραφικά άρθρα είναι μια σοβαρή ηθική παραβίαση που συμβαίνει όταν οι δημοσιογράφοι παρουσιάζουν τη δουλειά, τις ιδέες ή τα λόγια κάποιου άλλου ως δικά τους χωρίς να αποδίδουν τα κατάλληλα εύσημα, γεγονός που οδηγεί πολλές φορές σε διπλότυπα ειδησεογραφικών γεγονότων. Υπάρχουν διάφορες μορφές λογοκλοπής που μπορεί να εκδηλωθούν σε άρθρα ειδήσεων:

• Απευθείας αντιγραφή: Αντιγραφή αποσπασμάτων ή ολόκληρων τμημάτων ενός άρθρου ειδήσεων από άλλη πηγή χωρίς τη παροχή καμίας αναφοράς. Αυτό περιλαμβάνει αντιγραφή προτάσεων και παραγράφων λέξη προς λέξη.

• Παράφραση χωρίς αναφορά: Σε αυτόν τον τύπο λογοκλοπής, πραγματοποιείται αναδιατύπωση περιεχομένου από άλλη πηγή, αλλά χωρίς αναφορά του αρχικού συγγραφέα ή της δημοσίευσης. Αν και το κείμενο αναδιατυπώνεται, οι ιδέες και η δομή παραμένουν ουσιαστικά οι ίδιες με το αρχικό υλικό.

• Αυτολογοκλοπή: Οι δημοσιογράφοι μπορεί να είναι ένοχοι αυτολογοκλοπής όταν επαναχρησιμοποιούν το έργο τους που είχαν δημοσιέυσει στο παρελθόν χωρίς να υποδεικνύουν ότι δεν πρόκειται για εντελώς νέο περιεχόμενο.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Αυτό το φαινόμενο δημιουργεί σημαντικές ανησυχίες για τους καταναλωτές πληροφοριών, τους οργανισμούς ειδήσεων και την κοινωνία γενικότερα, υπογραμμίζοντας την ανάγκη για ισχυρούς μηχανισμούς ανίχνευσης για τη διατήρηση της ακεραιότητας, της αξιοπιστίας της διάδοσης των ειδήσεων.

Τα κείμενα μπορεί να είναι παρόμοια με δύο τρόπους λεξιλογικά και σημασιολογικά. Οι λέξεις που απαρτίζουν το κείμενο μοιάζουν λεξιλογικά αν έχουν παρόμοια ακολουθία χαρακτήρων. Η σημασιολογική ομοιότητα είναι μια μέτρηση που ορίζεται σε ένα σύνολο εγγράφων ή όρων, όπου η ιδέα της απόστασης μεταξύ των στοιχείων βασίζεται στην ομοιότητα του νοήματος ή του σημασιολογικού περιεχομένου τους σε αντίθεση με τη λεξιλογική ομοιότητα. Επιπλέον, πρέπει να σημειωθεί ότι η σημασία μιας λέξης συντίθεται μέσω των σχέσεών της με τα συμφραζόμενα. Συνεπώς, μια λεξική μονάδα μπορεί να φέρει ένα σημασιολογικό χαρακτηριστικό, το οποίο επιδρά και διαμορφώνει εν μέρει τη σημασία μιας γειτονικής λέξης [Δ18].

Δεδομένου ότι τα μοναδικά αναγνωριστικά εγγράφων δεν είναι δυνατά στις διάφορες πηγές, ο εντοπισμός διπλών πληροφοριών είναι απαραίτητος για την παραγωγή μη περιττών αποτελεσμάτων [CFGM02]. Η σημασία του εντοπισμού σχεδόν διπλότυπων ειδήσεων δεν μπορεί να υπερεκτιμηθεί, καθώς επηρεάζει άμεσα την ποιότητα και την αυθεντικότητα των πληροφοριών που είναι διαθέσιμες στο κοινό. Οι ειδησεογραφικοί οργανισμοί προσπαθούν να παρέχουν διαφορετικές προοπτικές και ολοκληρωμένη κάλυψη, αλλά ο πολλαπλασιασμός σχεδόν διπλών άρθρων μπορεί να θέσει σε κίνδυνο αυτές τις προσπάθειες. Οι αναγνώστες που συναντούν πολλά σχεδόν πανομοιότυπα άρθρα εκτίθενται σε περιττές πληροφορίες, μειώνοντας την ικανότητά τους να έχουν πρόσβαση σε νέες ιδέες ή αντίθετες απόψεις για ένα συγκεκριμένο ειδησεογραφικό γεγονός.

Από τεχνική άποψη, η ανίχνευση σχεδόν διπλότυπων ειδήσεων θέτει περίπλοκες προκλήσεις που διασταυρώνουν κλάδους όπως η ανάκτηση πληροφοριών, η επεξεργασία φυσικής γλώσσας και η μηχανική εκμάθηση. Ο εντοπισμός ομοιοτήτων και ανομοιοτήτων κειμένου σε ένα τεράστιο σύνολο άρθρων απαιτεί εξελιγμένους αλγόριθμους ικανούς να χειρίζονται τις αποχρώσεις της γλώσσας, τις συντακτικές δομές και το σημασιολογικό νόημα. Οι εξελίξεις σε αυτούς τους τομείς όχι μόνο ενισχύουν την αποτελεσματικότητα των σχεδόν διπλότυπων συστημάτων ανίχνευσης,

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αλλά έχουν επίσης ευρύτερες επιπτώσεις για την ανάκτηση πληροφοριών, τη σύσταση περιεχομένου και τη διαχείριση γνώσης σε διάφορους τομείς.

3.2 Τεχνικές ανίχνευσης διπλοτύπων και σχεδόν διπλοτύπων

Στα συστήματα βάσεων δεδομένων ένα ουσιαστικό βήμα για τον καθαρισμό δεδομένων και την ενοποίηση δεδομένων είναι η αναγνώριση των σχεδόν διπλότυπων εγγράφων (NDD) [AAE12]. Εμπνευσμένο από τις ανάγκες πολλών εργασιών του πραγματικού κόσμου κατά το χειρισμό μεγάλων συλλογών εγγράφων, η σχεδόν διπλότυπη ανίχνευση είναι ένα σημαντικό ερευνητικό πρόβλημα. Ο εντοπισμός σχεδόν διπλότυπου εγγράφου είναι το πρόβλημα της ταχείας εύρεσης όλων των εγγράφων των οποίων οι ομοιότητες είναι ίσες ή μεγαλύτερες από ένα δεδομένο όριο.

Σε πολλά σενάρια, δύο έγγραφα που δεν είναι ακριβώς πανομοιότυπα ενδέχεται να εξακολουθούν να περιέχουν το ίδιο περιεχόμενο και θα πρέπει να αντιμετωπίζονται ως διπλότυπα. Ωστόσο, το ποιο τμήμα των εγγράφων θα πρέπει να θεωρείται σημαντικό σε μια τέτοια σύγκριση εξαρτάται από την τελική εφαρμογή. Για παράδειγμα, οι ιστοσελίδες από διαφορετικούς ιστότοπους ενδέχεται να διαφέρουν μόνο στο πεδίο της κεφαλίδας ή υποσελίδας που υποδηλώνουν τη διεύθυνση URL του ιστότοπου και τον χρόνο ενημέρωσης.

Τα άρθρα ειδήσεων που εμφανίζονται σε διαφορετικές πύλες θα μπορούσαν να προέρχονται από την ίδια πηγή και επομένως να έχουν το ίδιο περιεχόμενο, αλλά μπορούν να αποδοθούν σε διαφορετικά πρότυπα ιστοτόπων με διαφημίσεις. Όταν ένα τμήμα ενός εγγράφου, όπως μια πρόταση ή μια παράγραφος, περιέχεται σε ένα άλλο έγγραφο, αυτά τα δύο έγγραφα θα μπορούσαν να θεωρηθούν σχεδόν διπλότυπα [HYK10].

Για την επίτευξη του εντοπισμού των διπλοτύπων, αρκετοί ερευνητές έχουν χρησιμοποιήσει διάφορες τεχνικές ανίχνευσης. Μερικές από αυτές είναι η χρήση μέτρων ομοιότητας, επώνυμων οντοτήτων, λέξεων-κλειδιών και μηχανικής μάθησης.

3.2.1 Μέτρα ομοιότητας

Από την άποψη της θεωρίας πληροφοριών, η ομοιότητα ορίζεται ως η κοινότητα μεταξύ δύο αποσπασμάτων κειμένου. Όσο μεγαλύτερη είναι η κοινότητα, τόσο μεγαλύτερη η ομοιότητα, και το αντίστροφο. Η ομοιότητα κειμένου γίνεται γρήγορα βασικό εργαλείο σε πολλές εργασίες που βασίζονται στην επεξεργασία φυσικής γλώσσας, όπως η αντιστοίχιση εγγράφων [WD20].

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Ο εντοπισμός σχεδόν διπλότυπων βοηθά στην ανάκτηση έγκυρων πληροφοριών με υψηλή ταχύτητα, μειώνοντας τα έξοδα αποθήκευσης. Η ποσοτικοποίηση δύο περιεχομένων κειμένου ως σχεδόν διπλότυπα μπορεί να γίνει με τη χρήση τεχνικών μέτρησης ομοιότητας κειμένου. Η μέτρηση της ομοιότητας είναι μια δημοφιλής μέθοδος για τον εντοπισμό του βαθμού συσχέτισης μεταξύ δύο κειμένων είτε με τη μορφή λέξης, πρότασης, παραγράφου ή εγγράφου [MPK20].

Οι τεχνικές αυτές, χρησιμοποιούν συνάρτηση ομοιότητας, η οποία μετρά τον βαθμό ομοιότητας μεταξύ δύο αντικειμένων και επιστρέφει μια τιμή στο διάστημα [0, 1]. Μια υψηλότερη τιμή ομοιότητας δείχνει ότι τα αντικείμενα είναι περισσότερο παρόμοια [CFGM02]. Έτσι, μπορούμε να αντιμετωπίσουμε ζεύγη αντικειμένων με υψηλή τιμή ομοιότητας ως σχεδόν διπλότυπα. [X+11].

Η ομοιότητα λέξεων εξετάζει όρους ή λέξεις ενώ η ομοιότητα εγγράφου εξετάζει έγγραφα που μπορεί να έχουν τη μορφή προτάσεων ή παραγράφων. Η ομοιότητα λεξιλογικού κειμένου αναλύει δύο κείμενα λαμβάνοντας υπόψη παραμέτρους όπως η σύνταξη, η δομή και το περιεχόμενο, ενώ η σημασιολογική ομοιότητα εξετάζει παραμέτρους όπως τη σημασιολογία, το νόημα και το πλαίσιο. Στη σημασιολογική ομοιότητα, η γραμματική και η αναγνώριση οντοτήτων παίζει σημαντικό ρόλο [MPK20].

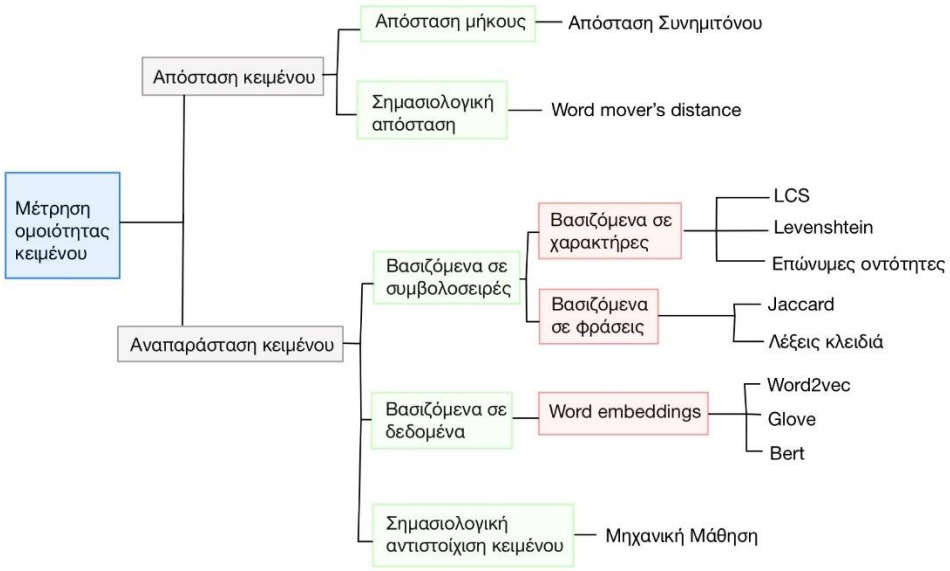
Οι σημασιολογικές σχέσεις μεταξύ των εγγράφων προσδιορίζουν παρόμοιες πληροφορίες περιεχομένου μεταξύ τους. Αυτό μπορεί να γίνει με την απόσταση μεταξύ της ορολογίας. Τα τελευταία χρόνια η επεξεργασία της φυσικής γλώσσας βοηθά στον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ των λέξεων και των προτάσεων με βάση τις συνυπάρχουσες λέξεις ή τη λεξιλογική βάση δεδομένων [DS22].

Η απόσταση κειμένου περιγράφει τη σημασιολογική εγγύτητα δύο λέξεων κειμένου από την προοπτική της απόστασης. Τεχνικές που χρησιμοποιούν μέτρα ομοιότητας υπολογίζουν μια τιμή ομοιότητας για κάθε ζεύγος εγγράφων και για να κατανοήσουμε ότι ένα έγγραφο είναι παρόμοιο με ένα άλλο, η τιμή ομοιότητάς του πρέπει να υπερβαίνει κάποια τιμή κατωφλίου. Σε προσεγγίσεις που χρησιμοποιούν μέτρα ομοιότητας, η τιμή που σχετίζεται με το όριο είναι πολύ σημαντική.

Ο καθορισμός μιας υψηλής τιμής για το όριο θα έχει ως αποτέλεσμα την απώλεια εγγράφων που είναι πραγματικά διπλότυπα. Έχουν γίνει αρκετές προσπάθειες

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

από ερευνητές για τον προσδιορισμό της ομοιότητας ενός εγγράφου με άλλο έγγραφο. [U09]. Παρακάτω αναλύονται τα κυριότερα μέτρα ομοιότητας.



Εικόνα 3-1: Διαγραμματική απεικόνιση των μέτρων ομοιότητας

3.2.1.1 Ομοιότητα συνημίτονου (Cosine Similarity)

Η ομοιότητα συνημίτονου είναι ένα ευρέως χρησιμοποιούμενο μέτρο στην ανάλυση ομοιότητας κειμένου. Κατά τη μέτρηση της απόστασης συνημίτονου, αντί να μετρηθεί η απόσταση μεταξύ των δύο σημείων, μετριέται η γωνία που αντιστοιχεί στα δύο σημεία του διανυσματικού χώρου. Η ομοιότητα υπολογίζεται μετρώντας το συνημίτονο της γωνίας μεταξύ των δύο αυτών διανυσμάτων [MPK20]. Πιο αναλυτικά, η ομοιότητα συνημίτονου είναι η γωνία μεταξύ δύο διανυσμάτων εγγράφου σε n διαστατικό χώρο. Λαμβάνοντας υπόψη δύο διανύσματα εγγράφου Sa και Sb, η ομοιότητα συνημίτονου, θ, αναπαρίσταται ως:

$$\mathrm{Sim}(S_a, S_b) = \cos\Theta = \frac{S_a \cdot S_b}{||S_a|| \cdot ||S_b||}$$

Εικόνα 3-2: Τύπος απόστασης συνημίτονου

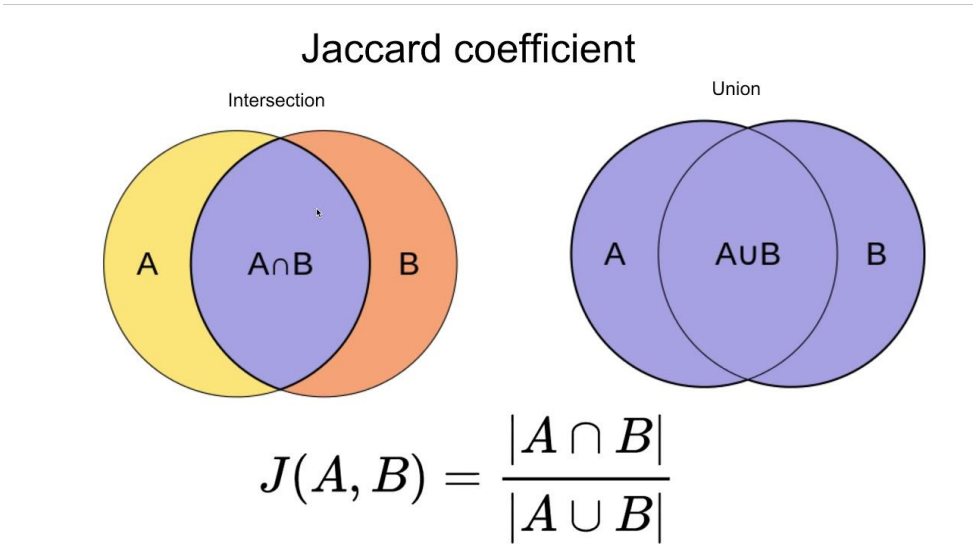
Συγκρίνοντας τις διανυσματικές αναπαραστάσεις των εγγράφων, η ομοιότητα συνημίτονου καταγράφει την ομοιότητα όσον αφορά τους προσανατολισμούς τους και όχι τα μεγέθη τους. Σύμφωνα με τη βιβλιογραφία, παρατηρήθηκε πως το μέτρο συνημίτονου είναι ανεξάρτητο από τον αριθμό των εμφανίσεων μιας λέξης σε ένα έγγραφο [U09]. Αυτό το μέτρο είναι ιδιαίτερα αποτελεσματικό για την καταγραφή της

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

σημασιολογικής ομοιότητας [DS22] και χρησιμοποιείται συχνά σε εργασίες όπως η ομαδοποίηση εγγράφων και η ανάκτηση πληροφοριών.

3.2.1.2 Ομοιότητα Jaccard (Jaccard Similarity)

Η ομοιότητα Jaccard μετρά την ομοιότητα μεταξύ δύο συνόλων υπολογίζοντας το μέγεθος της τομής τους διαιρούμενο με το μέγεθος της ένωσής τους.



Εικόνα 3-3: Οπτικοποίηση του τύπου Jaccard με διαγράμματα Venn. Πηγή: medium.com

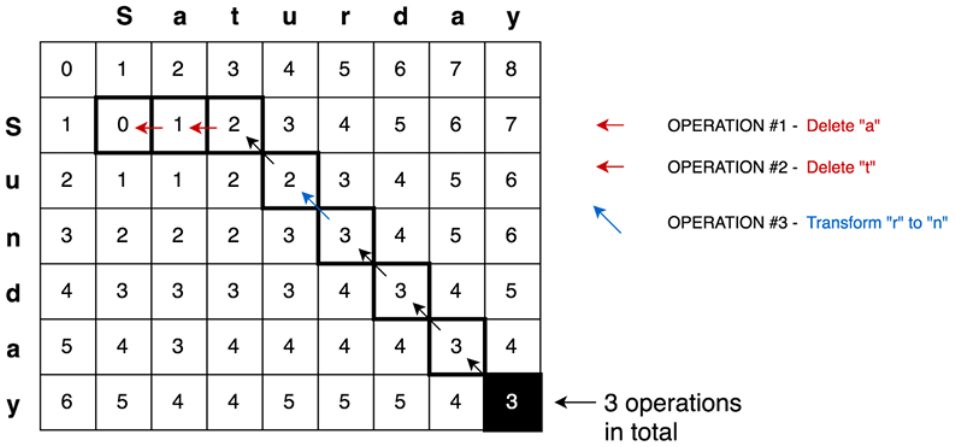
Στην ανίχνευση αντιγραφής κειμένου, τα έγγραφα αντιπροσωπεύονται ως σύνολα λέξεων ή n-grams. Η ομοιότητα Jaccard είναι ιδιαίτερα χρήσιμη για τη μέτρηση της επικάλυψης μεταξύ συνόλων όρων, παρέχοντας ένα μέτρο ομοιότητας που είναι ανθεκτικό στις διακυμάνσεις του μήκους του εγγράφου [MPK20].

3.2.1.3 Απόσταση Levenshtein (Edit Distance / Levenshtein Distance)

Η επεξεργασία απόστασης, επίσης γνωστή ως απόσταση Levenshtein, ποσοτικοποιεί τον ελάχιστο αριθμό πράξεων όπως εισαγωγές, διαγραφές και αντικαταστάσεις που απαιτούνται για τη μετατροπή μιας συμβολοσειράς σε άλλη [X+11]. Στο πλαίσιο της ανίχνευσης διπλότυπου κειμένου, η απόσταση επεξεργασίας μπορεί να χρησιμοποιηθεί για τη μέτρηση της ομοιότητας μεταξύ δύο κειμένων, αξιολογώντας τον ελάχιστο αριθμό πράξεων που απαιτούνται για τη μετατροπή του ενός κειμένου στο άλλο. Αυτό το μέτρο είναι χρήσιμο για τον εντοπισμό σχεδόν διπλότυπων περιπτώσεων με μικρές παραλλαγές ή τροποποιήσεις κειμένου. Παραδοσιακά, η ομοιότητα κειμένου αξιολογείται με τη μέτρηση της απόστασης, η οποία χρησιμοποιεί τα αριθμητικά χαρακτηριστικά του κειμένου για τον υπολογισμό της απόστασης του διανυσματικού κειμένου [WD20].

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

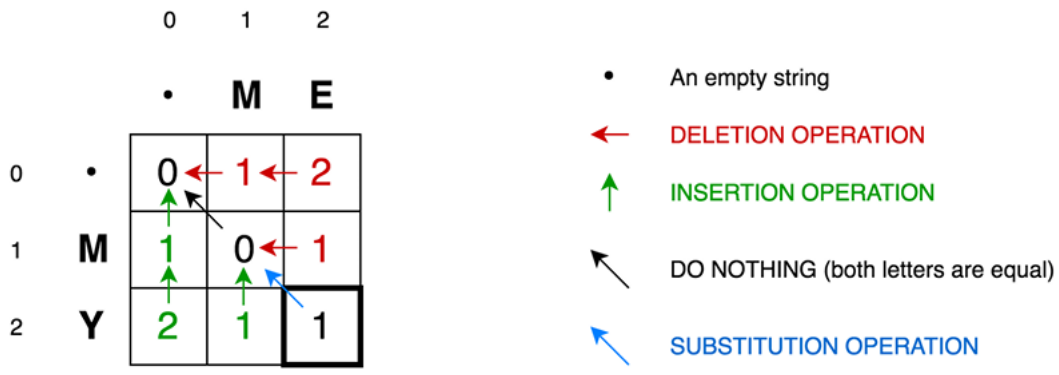
Ένας πίνακας αρχικοποιείται για να μετρήσει την απόσταση Levenshtein μεταξύ του πρώτου χαρακτήρα μιας λέξης και του τελευταίου χαρακτήρα της άλλης λέξης. Η μήτρα μπορεί να γεμιστεί από την επάνω αριστερή έως την κάτω δεξιά γωνία. Κάθε άλμα στο κείμενο αντιστοιχεί σε μια εισαγωγή ή μια διαγραφή που αποφασίζεται στο δεύτερο βήμα. Το κόστος κάθε λειτουργίας ορίζεται συνήθως σε 1. Το διαγώνιο άλμα κοστίζει είτε ένα είτε μηδενικό ανάλογα με το αν ταιριάζουν οι δύο χαρακτήρες στη γραμμή και στη στήλη. Κάθε κελί προσπαθεί πάντα να ελαχιστοποιήσει το κόστος τοπικά.



Εικόνα 3-4: Παράδειγμα μεθόδου Levenshtein. Πηγή: [PR]

Στο πρώτο βήμα, πρέπει να δημιουργηθεί ένας πίνακας με τον οποίο οι λέξεις τοποθετούνται στις σειρές και τις στήλες ανεξάρτητα από το αν ο αριθμός των γραμμάτων είναι ίδιος ή διαφορετικός. Αρχικά, τοποθετούνται οι αρχικές τιμές στις λέξεις σύμφωνα με τη σειρά τους από την αρχή μέχρι το τέλος. Μετά από κάθε γράμμα, αυξάνεται η τιμή κατά ένα. Και στις δύο λέξεις τα κελιά δίπλα τους είναι γεμάτα με τιμές σε αύξουσα σειρά. Για κάθε κελί, εάν τα συγκριτικά γράμματα είναι ίσα, τότε πρέπει να εκχωρηθεί απευθείας η προηγούμενη τιμή διαγώνιου κελιού του τρέχοντος κελιού. Διαφορετικά, εάν τα συγκριτικά γράμματα δεν είναι ίδια μεταξύ τους, τότε αυξάνονται οι τρεις τιμές γύρω από αυτό - αριστερά, επάνω και διαγώνια - πάνω αριστερά. Μετά την αύξηση, η μικρότερη τιμή αυτών των αποτελεσμάτων επιλέγεται ως η νέα τιμή του τρέχοντος κελιού.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Εικόνα 3-5: Τρόπος λειτουργίας της μεθόδου Levensthein. Πηγή: [PR]

Στον πίνακα που δημιουργείται, δίνεται έμφαση στο τελευταίο στοιχείο του πίνακα που βρίσκεται στο δεξιότερο και χαμηλότερο κελί σε ολόκληρο το πλέγμα. Μετά από αυτό το επιλεγόμενο κελί, ελέγχονται τα τρία κελιά γύρω από το τρέχον κελί - επάνω, αριστερό και αριστερό επάνω (διαγώνιο) - και επιλέγεται το ελάχιστο των τριών τιμών των κελιών ως το κελί-στόχος. Πριν επιλεγεί εκείνο το κελί, πρέπει να ληφθεί μια απόφαση για το ποια πράξη πρέπει να γίνει.

Επιπλέον, εάν επιλεγεί η αντικατάσταση ως η βέλτιστη λειτουργία, πρέπει να αποφασίσουμε ποια γράμματα πρέπει να ανταλλάσσονται μεταξύ τους. Εάν η ελάχιστη τιμή βρίσκεται στο αριστερό κελί, μπορούμε να συμπεράνουμε ότι η λειτουργία μας θα πρέπει να τακτοποιηθεί ως το τρέχον γράμμα διαγραφής που ελέγχεται έναντι άλλων. Εάν η ελάχιστη τιμή βρίσκεται σε ένα επάνω κελί, μπορούμε να συμπεράνουμε ότι η απαραίτητη λειτουργία για την κατάλληλη λύση είναι η εισαγωγή. Τέλος, εάν το διαγώνιο κελί έχει τη μικρότερη τιμή, τότε αυτό σημαίνει ότι μια λειτουργία αντικατάστασης θα πρέπει να συμπεριληφθεί στη λίστα των απαιτούμενων πράξεων. Εάν δεν βρεθεί ελάχιστη τιμή, δηλαδή όλα τα κελιά είναι ίσα, τότε παραλείπουμε αυτό το σημείο χωρίς να κάνουμε τίποτα επειδή είναι ακριβώς τα ίδια γράμματα [PR].

3.2.1.4 Μεγαλύτερη κοινή υποακολουθία (Longest Common Subsequence (LCS))

Η μεγαλύτερη κοινή υποακολουθία είναι ιδιαίτερα αποτελεσματική για την καταγραφή δομικών ομοιοτήτων και τον εντοπισμό σχεδόν διπλότυπων περιπτώσεων με αναδιάταξη περιεχομένου. Είναι μια κλασική αλγοριθμική έννοια που χρησιμοποιείται για την εύρεση της μεγαλύτερης κοινής υποακολουθίας σε δύο ή περισσότερες ακολουθίες. Στο πλαίσιο της ανίχνευσης διπλότυπου κειμένου, χρησιμοποιείται για τον εντοπισμό της μεγαλύτερης κοινής υποακολουθίας μεταξύ δύο εγγράφων κειμένου.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

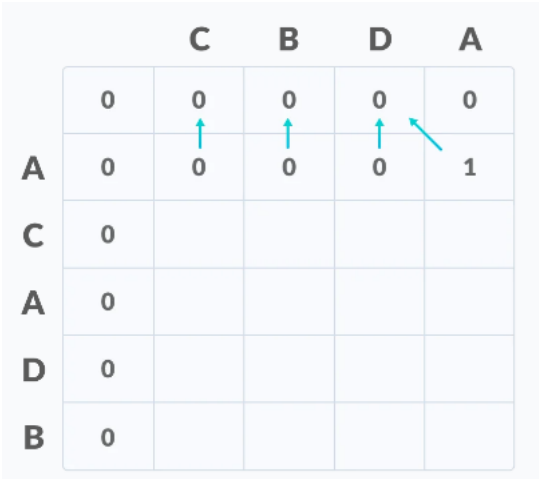
Γενικά, μια υποακολουθία είναι μια ακολουθία που μπορεί να προέλθει από μια άλλη ακολουθία διαγράφοντας ορισμένα στοιχεία ή κανένα στοιχείο χωρίς να αλλάξει η σειρά των υπολοίπων στοιχείων. Η μεγαλύτερη κοινή υποακολουθία μεταξύ δύο αλληλουχιών είναι η μεγαλύτερη υποακολουθία που είναι κοινή και στις δύο αλληλουχίες.

Ο αλγόριθμος LCS στοχεύει να βρει τη μεγαλύτερη κοινή υποακολουθία συγκρίνοντας επαναληπτικά στοιχεία μεταξύ των ακολουθιών. Συνήθως χρησιμοποιεί τεχνικές δυναμικού προγραμματισμού για τον αποτελεσματικό υπολογισμό του LCS. Ο αλγόριθμος κατασκευάζει έναν πίνακα ή έναν πίνακα όπου κάθε κελί αποθηκεύει το μήκος του LCS μέχρι εκείνο το σημείο. Λαμβάνοντας υπόψη τα στοιχεία των ακολουθιών και ενημερώνοντας τις τιμές στη μήτρα με βάση τις αντιστοιχίσεις ή τις αναντιστοιχίες, ο αλγόριθμος δημιουργεί σταδιακά το LCS. Το μήκος του LCS υποδεικνύει τον βαθμό ομοιότητας μεταξύ των κειμένων. Τα μεγαλύτερα μήκη LCS υποδηλώνουν μεγαλύτερη επικάλυψη κειμένου, ενώ τα μικρότερα μήκη συνεπάγονται περισσότερο αποκλίνοντα ή ανόμοιο περιεχόμενο.

Αναλυτικότερα, παρακάτω αναγράφονται τα βήματα για την εύρεση της μεγαλύτερης κοινής υποακολουθίας.

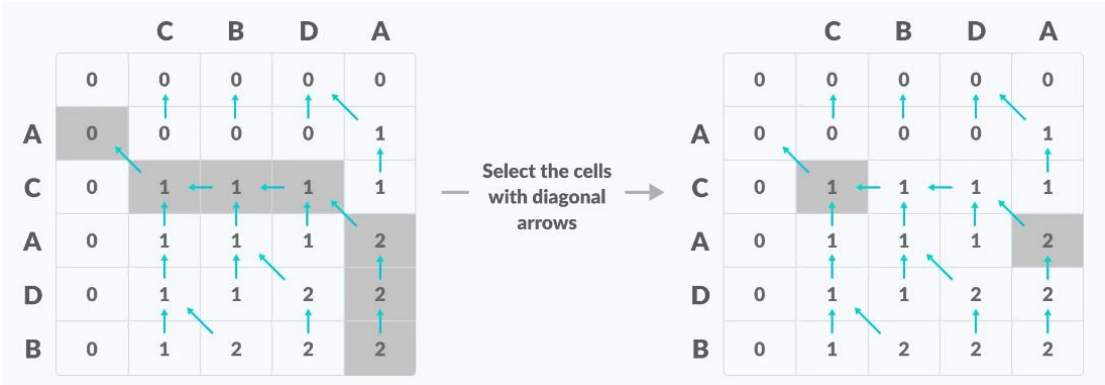
1. Δημιουργία ενός πίνακα διαστάσεων n+1*m+1 όπου n και m είναι τα μήκη των X και Y ακολουθιών αντίστοιχα. Η πρώτη σειρά και η πρώτη στήλη γεμίζουν με μηδενικά.
2. Συμπλήρωση κάθε κελιού του πίνακα χρησιμοποιώντας την παρακάτω λογική.
    a) Εάν ο χαρακτήρας που αντιστοιχεί στην τρέχουσα σειρά και την τρέχουσα στήλη ταιριάζουν, γίνεται συμπλήρωση του τρέχοντος κελιού προσθέτοντας ένα στο διαγώνιο στοιχείο. Επιπλέον, τοποθετείται ένα βέλος στο διαγώνιο κελί.
    b) Διαφορετικά, η μέγιστη τιμή από την προηγούμενη στήλη και το στοιχείο της προηγούμενης γραμμής χρησιμοποιούνται για τη συμπλήρωση του τρέχοντος κελιού. Τοποθετείται ένα βέλος στο κελί με τη μέγιστη τιμή. Εάν είναι ίσα, επιλέγεται κάποιο από αυτά.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Εικόνα 3-6: Γέμισμα μήτρας LCS. Πηγή: programiz.com

3. Το βήμα 2 επαναλαμβάνεται μέχρι να γεμίσει ο πίνακας.
4. Η τιμή στην τελευταία σειρά και στην τελευταία στήλη είναι το μήκος της μεγαλύτερης κοινής υποακολουθίας.
5. Για να βρεθεί η μεγαλύτερη κοινή υποακολουθία, γίνεται εκκίνηση από το τελευταίο στοιχείο και ακολουθείται η κατεύθυνση του βέλους. Τα στοιχεία που αντιστοιχούν στο διαγώνιο βέλος σχηματίζουν τελικά τη μεγαλύτερη κοινή υποακολουθία.



Εικόνα 3-7: Τελική μήτρα του αλγορίθμου LCS. Πηγή: programiz.com

Παρόλο που το LCS είναι ένα πολύτιμο εργαλείο για τον εντοπισμό κοινών υποακολουθιών, δεν λαμβάνει υπόψη τη συγκεκριμένη σειρά ή θέση των στοιχείων εντός της υποακολουθίας. Αυτό σημαίνει ότι το LCS μπορεί να ανιχνεύσει σχεδόν διπλότυπα στιγμιότυπα ακόμα και όταν το περιεχόμενο έχει αναδιατυπωθεί ή αναδιαταχθεί σημαντικά. Επιπλέον, το LCS δεν καταγράφει σημασιολογικές ομοιότητες ή διαφορές μεταξύ των κειμένων, εστιάζοντας αποκλειστικά στο μήκος της

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

κοινής υποακολουθίας. Ως αποτέλεσμα, ενδέχεται να μην είναι κατάλληλο για τον εντοπισμό σχεδόν διπλότυπων περιπτώσεων με ανεπαίσθητες παραλλαγές ή σημασιολογικές αλλαγές.

Τα μέτρα ομοιότητας "Απόσταση Levenshtein" και "Μεγαλύτερη κοινή υποακολουθία" που βασίζονται στη μέτρηση απόστασης καταγράφουν την ομοιότητα μεταξύ δύο προτάσεων ή παραγράφων μόνο εάν έχουν κοινές λέξεις ή όρους. Επιπλέον, δεν μπορούν να χειριστούν συνώνυμες λέξεις. Για παράδειγμα, λέξεις όπως «χαρά» και «ευτυχισμένος» που έχουν παρόμοια σημασία, θα αντιμετωπίζονται ως διαφορετικές λέξεις από αυτές τις μετρήσεις. Έτσι, για να είναι εφικτός ο εντοπισμός τέτοιων λέξεων, τα «word embeddings» μπορούν να παίξουν μεγάλο ρόλο στην ανάλυση ομοιότητας [MPK20].

3.2.1.4.1. Ο αλγόριθμος Ratcliff/Obershelp

Ο αλγόριθμος Ratcliff/Obershelp, γνωστός και ως αλγόριθμος αναγνώρισης προτύπων Ratcliff/Obershelp, είναι ένας αλγόριθμος αντιστοίχισης συμβολοσειρών που χρησιμοποιείται για την εύρεση της ομοιότητας μεταξύ δύο συμβολοσειρών. Αναπτύχθηκε από τους John Ratcliff και D. Wayne Obershelp και είναι ιδιαίτερα χρήσιμος για τον εντοπισμό της μεγαλύτερης συνεχόμενης αλληλουχίας αντιστοίχισης μεταξύ δύο συμβολοσειρών.

Λειτουργεί με την αναδρομική αναζήτηση για τη μεγαλύτερη κοινή υποακολουθία μεταξύ δύο συμβολοειρών. Το σχήμα εργασίας του αλγόριθμου Ratcliff/Obershelp [KR20] είναι το εξής:

- Πρώτον, συγκρίνονται δύο συμβολοσειρές και εντοπίζεται η μεγαλύτερη κοινή υποακολουθία μεταξύ τους. Αυτή η ακολουθία επισημαίνεται ως σημείο ελέγχου.
- Στη συνέχεια, ο αλγόριθμος εφαρμόζεται αναδρομικά στα μέρη των συμβολοσειρών που βρίσκονται σε κάθε πλευρά της αντιστοίχισης. Ουσιαστικά επαναλαμβάνει τη διαδικασία με τις υποσυμβολοσειρές και στις δύο πλευρές της μεγαλύτερης αντίστοιχης υποσυμβολοσειράς.
- Η αναδρομή συνεχίζεται έως ότου δεν υπάρχουν άλλες κοινές υποσυμβολοσειρές που να ταιριάζουν.

Η παράμετρος επιστροφής υπολογίζεται από αυτόν τον τύπο:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

$$D_{ro} = \frac{2 * K_m}{|s_1| + |s_2|}$$

Εικόνα 3-8: Τύπος αλγορίθμου Ratcliff/Obershelp

όπου Km είναι ένας αριθμός παρόμοιων χαρακτήρων και |s1| και |s2| είναι τα μήκη των δύο συμβολοσειρών εισόδου που συγκρίνονται, συχνά αναφερόμενες ως "string1" και "string2". Η τιμή που δίνεται ως αποτέλεσμα του υπολογισμού στη συνέχεια χρησιμοποιείται ως ποσοστό για να δείξει πόσο μοιάζουν δύο συμβολοσειρές.

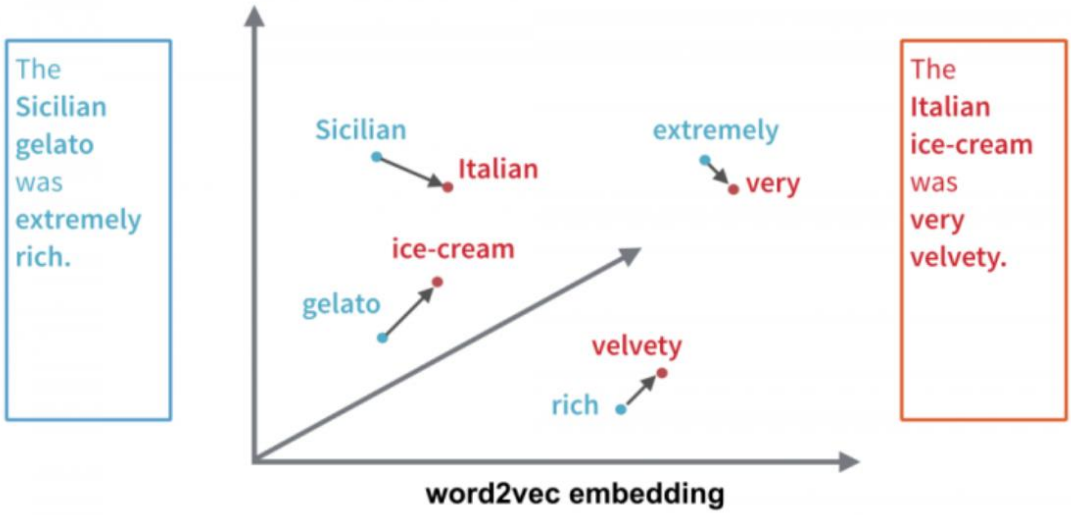
Ο αλγόριθμος Ratcliff/Obershelp είναι σχετικά απλός, αλλά μπορεί να είναι πολύ αποτελεσματικός στον εντοπισμό ομοιοτήτων μεταξύ των συμβολοσειρών, ειδικά όταν υπάρχει κάποιο επίπεδο θορύβου ή διακύμανση μεταξύ τους. Χρησιμοποιείται σε εφαρμογές όπως η σύγκριση κειμένου, η αφαίρεση των διπλοτύπων και ο εντοπισμός λογοκλοπής, όπου η εύρεση παρόμοιου περιεχομένου είναι απαραίτητη.

3.2.1.5 Word Mover's Distance (WMD)

Το WMD είναι ένα μέτρο που ποσοτικοποιεί την ανομοιότητα μεταξύ δύο κειμένων με βάση τη "μετακίνηση" των word embeddings από το ένα κείμενο στο άλλο. Τα word embeddings είναι διανυσματικές αναπαραστάσεις λέξεων που αποτυπώνουν τις σημασιολογικές έννοιες και τις σχέσεις τους με τα συμφραζόμενα. Τα δημοφιλή μοντέλα ενσωμάτωσης λέξεων όπως το Word2Vec, το GloVe ή το FastText μαθαίνουν κατανεμημένες αναπαραστάσεις όπου παρόμοιες λέξεις βρίσκονται πιο κοντά η μία στην άλλη στον διανυσματικό χώρο. Αυτές οι ενσωματώσεις χρησιμεύουν ως βάση για τον υπολογισμό των σημασιολογικών ομοιοτήτων μεταξύ των λέξεων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Το WMD υπολογίζει την ελάχιστη απόσταση που απαιτείται για τη μεταφορά των word embeddings από το ένα έγγραφο στο άλλο, λαμβάνοντας υπόψη τις σημασιολογικές ομοιότητες μεταξύ των λέξεων. Είναι ένα ισχυρό μέτρο που καταγράφει τη σημασιολογική σημασία και το πλαίσιο των λέξεων, καθιστώντας το κατάλληλο για τον εντοπισμό σχεδόν διπλών περιπτώσεων που παρουσιάζουν λεπτές παραλλαγές.



Εικόνα 3-9: Μέθοδος Word Mover's Distance. Πηγή: tech.opentable.com

Η παραπάνω εικόνα δείχνει σχηματικά πώς κάθε λέξη στο πρώτο έγγραφο βρίσκεται σε σχέση με τις πλησιέστερες λέξεις στο άλλο έγγραφο εντός του χώρου ενσωμάτωσης word2vec. Η απόσταση μεταξύ των δύο εγγράφων μπορεί να υπολογιστεί ως η ελάχιστη αθροιστική απόσταση που πρέπει να διανύσουν οι λέξεις από το πρώτο έγγραφο για να ταιριάζουν ακριβώς με το δεύτερο έγγραφο.

Το WMD μπορεί να είναι υπολογιστικά ακριβό, ειδικά όταν έχουμε να κάνουμε με μεγάλα έγγραφα ή εκτεταμένο λεξιλόγιο. Ο υπολογισμός περιλαμβάνει σύγκριση word embeddings και επίλυση προβλημάτων βελτιστοποίησης, τα οποία μπορεί να είναι χρονοβόρα. Επιπλέον, το WMD ενδέχεται να μην είναι κατάλληλο για όλους τους τύπους εργασιών ανίχνευσης αντιγραφής κειμένου, ιδιαίτερα όταν η εστίαση είναι στην καταγραφή δομικών μοτίβων, επικάλυψης κειμένου ή συγκεκριμένων γλωσσικών χαρακτηριστικών.

3.2.1.6 Επίλογος μέτρων ομοιότητας

Αυτά τα μέτρα ομοιότητας χρησιμοποιούνται συχνά μεμονωμένα ή σε συνδυασμό [AAE12], ανάλογα με τις ειδικές απαιτήσεις της εργασίας ανίχνευσης

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αντιγραφής κειμένου. Η επιλογή του μέτρου εξαρτάται από τα χαρακτηριστικά των δεδομένων, το επιθυμητό επίπεδο ευκρίνειας και τους συγκεκριμένους στόχους του συστήματος ανίχνευσης διπλοτύπων.

Ένα χαρακτηριστικό παράδειγμα είναι ο υπολογισμός ομοιότητας κατά ζεύγη που ασχολείται με την εύρεση ζευγών αντικειμένων σε ένα μεγάλο σύνολο δεδομένων παρόμοιων σύμφωνα με κάποιο μέτρο. Αυτό το πρόβλημα συναντάται συχνά σε εφαρμογές επεξεργασίας κειμένου, για παράδειγμα, ομαδοποίηση για μάθηση χωρίς επίβλεψη. Το σχεδόν διπλότυπο έγγραφο ολοκληρώνει τις συγκρίσεις ομοιότητας ανά ζεύγη σε δύο βήματα, όπου είναι η δημιουργία ανεστραμμένου ευρετηρίου και στη συνέχεια υπολογισμοί ομοιότητας με αυτό.

Αξιοποιώντας αυτά τα μέτρα ομοιότητας, οι ερευνητές μπορούν να αξιολογήσουν αποτελεσματικά τον βαθμό ομοιότητας μεταξύ εγγράφων κειμένου και να εντοπίσουν σχεδόν διπλότυπες περιπτώσεις με διαφορετικά επίπεδα επικάλυψης κειμένου.

3.2.2 Επώνυμες οντότητες

Οι επώνυμες οντότητες διαδραματίζουν κρίσιμο ρόλο στον εντοπισμό διπλότυπου κειμένου, παρέχοντας πολύτιμες πληροφορίες για τον εντοπισμό σχεδόν διπλότυπων περιπτώσεων. Οι επώνυμες οντότητες αναφέρονται σε συγκεκριμένες οντότητες, όπως πρόσωπα, οργανισμούς, τοποθεσίες, ημερομηνίες και άλλες ονομαστικές έννοιες που αναφέρονται σε κάποιο κείμενο. Ενσωματώνουν σημαντικές σημασιολογικές πληροφορίες που βοηθούν στην κατανόηση του περιεχομένου, του πλαισίου και της δομής των εγγράφων κειμένου. Η Αναγνώριση Επώνυμων Οντοτήτων (Named Entity Recognition) αναπτύσσεται ως δευτερεύων κομμάτι της εξαγωγής πληροφοριών, επειδή έγινε αντιληπτό οι μονάδες πληροφοριών που αναφέρθηκαν παραπάνω, είναι τα βασικά σημεία για την εξαγωγή χρήσιμων δεδομένων [U09].

Οι επώνυμες οντότητες μπορούν να χρησιμεύσουν ως διακριτικά χαρακτηριστικά για τη διάκριση μεταξύ διαφορετικών εγγράφων. Συγκρίνοντας την παρουσία, την απουσία ή τις παραλλαγές σε επώνυμες οντότητες, τα μέτρα ομοιότητας μπορούν να καταγράψουν τον βαθμό επικάλυψης ή απόκλισης μεταξύ των κειμένων. Για παράδειγμα, εάν δύο έγγραφα έχουν πανομοιότυπες οντότητες με βάση το όνομα, μπορεί να υποδηλώνει μεγαλύτερη πιθανότητα σχεδόν διπλότυπου κειμένου.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Επιπλέον, οι επώνυμες οντότητες παρέχουν πληροφορίες σχετικά με τα συμφραζόμενα που μπορούν να βοηθήσουν στον εντοπισμό σχεδόν διπλότυπων περιπτώσεων. Για παράδειγμα, η αναφορά συγκεκριμένων ατόμων, οργανισμών ή τοποθεσιών σε άρθρα ειδήσεων μπορεί να βοηθήσει στον εντοπισμό διπλότυπου περιεχομένου που περιστρέφεται γύρω από τα ίδια γεγονότα ή ειδήσεις, που μοιράζονται κοινές οντότητες αλλά μπορεί να παρουσιάζουν παραλλαγές σε άλλες πτυχές κειμένου.

Στη γλωσσική προεπεξεργασία δημιουργούνται πιθανές ονομαστικές οντότητες χρησιμοποιώντας σημεία στίξης ή κεφαλαία. Επίσης, σε αυτό το στάδιο προσδιορίζονται οντότητες που αποτελούνται από περισσότερες από μία λέξεις. Παρόλα αυτά, υπάρχουν ορισμένα προβλήματα σχετικά με τα κεφαλαία γράμματα. Τα κεφαλαία γράμματα είναι το σημείο εκκίνησης για την ανίχνευση επώνυμων οντοτήτων, αλλά κάθε λέξη που ξεκινά με κεφαλαίο γράμμα μπορεί να μην είναι σωστό

Τοποθεσία Πρόσωπο Ημερομηνία

Επισήμανε σήμερα το πρωί, ο Έλληνας πρωθυπουργός, Κυριάκος Μητσοτάκης, ότι θα είναι η πρώτη συνάντησή του με τον Τούρκο πρόεδρο μετά τις πρόσφατες εκλογές σε Ελλάδα και Τουρκία.

Εικόνα 3-10: Παράδειγμα Αναγνώρισης Επώνυμων Οντοτήτων

ουσιαστικό. Για παράδειγμα, στην πρόταση «Επισήμανε σήμερα το πρωί, ο Έλληνας πρωθυπουργός, Κυριάκος Μητσοτάκης, ότι θα είναι η πρώτη συνάντησή του με τον Τούρκο πρόεδρο μετά τις πρόσφατες εκλογές σε Ελλάδα και Τουρκία.», το πρώτο γράμμα της λέξης «Επισήμανε» είναι κεφαλαίο, επειδή βρίσκεται στην αρχή μιας πρότασης, αλλά το σύστημα δεν μπορεί να κατανοήσει αυτήν την περίπτωση και προσδιορίζει το "Επισήμανε" ως οντότητα, αν και δεν είναι [U09].

Συνοπτικά, οι επώνυμες οντότητες χρησιμεύουν ως πολύτιμοι δείκτες για τον εντοπισμό διπλότυπου κειμένου. Αξιοποιώντας την παρουσία και τις παραλλαγές επώνυμων οντοτήτων, μπορεί κανείς να εντοπίσει αποτελεσματικά σχεδόν διπλότυπα στιγμιότυπα που μοιράζονται παρόμοιες οντότητες ή περιστρέφονται γύρω από κοινά θέματα. Η ενσωμάτωση επώνυμων οντοτήτων στην ανάλυση κειμένου, ενισχύει την ακρίβεια και την ευαισθησία των συστημάτων ανίχνευσης αντιγραφής κειμένου, επιτρέποντας τη βαθύτερη κατανόηση των σημασιολογικών σχέσεων και των δομικών προτύπων μεταξύ των εγγράφων κειμένου.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

3.2.3 Λέξεις κλειδιά

Οι λέξεις-κλειδιά, παρόμοιες με τις επώνυμες οντότητες, είναι σημαντικά στοιχεία για τον εντοπισμό διπλότυπου κειμένου, συμβάλλοντας στον εντοπισμό και την ανάλυση σχεδόν διπλότυπων περιπτώσεων. Οι λέξεις-κλειδιά αναφέρονται σε σημαντικούς όρους ή φράσεις που ενσωματώνουν την ουσία και την επίκαιρη εστίαση ενός εγγράφου. Παρέχουν βασικές γνώσεις για το περιεχόμενο και διευκολύνουν τη σύγκριση και την κατηγοριοποίηση των κειμένων εγγράφων.

Συγκεκριμένα, οι λέξεις-κλειδιά χρησιμεύουν ως μια συνοπτική αναπαράσταση του περιεχομένου ενός εγγράφου. Αποτυπώνουν τα κύρια θέματα, τις έννοιες και τις ιδέες που αναφέρονται στο κείμενο. Συγκρίνοντας τις λέξεις-κλειδιά διαφορετικών εγγράφων, μπορεί κανείς να εκτιμήσει την ομοιότητα ή την ανομοιότητα ως προς τα εξέχοντα θέματα ή το θέμα. Παρόμοια σύνολα λέξεων-κλειδιών μπορεί να υποδηλώνουν υψηλότερη πιθανότητα σχεδόν διπλότυπου κειμένου.

Ακόμη, μπορούν να χρησιμοποιηθούν ως διακριτικά χαρακτηριστικά για τη διάκριση μεταξύ διαφορετικών εγγράφων. Συγκρίνοντας την παρουσία, την απουσία ή τις παραλλαγές των λέξεων-κλειδιών, τα μέτρα ομοιότητας μπορούν να προσδιορίσουν την έκταση της επικάλυψης ή της απόκλισης μεταξύ των κειμένων. Η συνεπής χρήση λέξεων-κλειδιών ή συγκεκριμένα μοτίβα λέξεων-κλειδιών μπορεί να υποδηλώνουν πιθανές σχεδόν διπλότυπες περιπτώσεις.

Οι λέξεις-κλειδιά παρέχουν πληροφορίες σχετικά με τα συμφραζόμενα που βοηθούν στον εντοπισμό σχεδόν διπλότυπων περιπτώσεων. Λαμβάνοντας υπόψη την παρουσία συγκεκριμένων λέξεων-κλειδιών, ιδιαίτερα εκείνων που σχετίζονται με γεγονότα, τοποθεσίες ή εξειδικευμένους τομείς, καθίσταται δυνατός ο εντοπισμός περιεχομένου που περιστρέφεται γύρω από παρόμοια θέματα ή αφηγήσεις. Αυτή η συμφραζόμενη σημασία συμβάλλει στον εντοπισμό και την ομαδοποίηση σχεδόν διπλότυπων περιπτώσεων.

Επιπρόσθετα, μπορούν να αποκαλύψουν δομικά μοτίβα και ευθυγραμμίσεις μεταξύ εγγράφων. Εξετάζοντας τη διάταξη και την κατανομή των λέξεων-κλειδιών μέσα στα κείμενα, καθίσταται δυνατός ο εντοπισμός σχεδόν διπλότυπων περιπτώσεων με αναδιατυπωμένο ή αναδιαταγμένο περιεχόμενο. Η ανάλυση της συνέπειας ή της παραλλαγής στην τοποθέτηση λέξεων-κλειδιών βοηθά στην αποτύπωση διαφοροποιημένων ομοιοτήτων και διαφορών μεταξύ των εγγράφων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Αξίζει να σημειωθεί, πως οι λέξεις-κλειδιά μπορούν να αξιοποιηθούν ως χαρακτηριστικά σε μοντέλα μηχανικής εκμάθησης ή σε αλγόριθμους ομοιότητας. Με την αναπαράσταση εγγράφων μέσω των διανομών λέξεων-κλειδιών τους, μπορούν να εξαχθούν σημαντικά χαρακτηριστικά που αποτυπώνουν την ουσία του περιεχομένου. Αυτά τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για εκπαίδευση συστημάτων κατηγοριοποίησης ή μετρήσεων ομοιότητας, επιτρέποντας την αυτοματοποιημένη ανίχνευση σχεδόν διπλότυπων περιπτώσεων.

Επιπλέον, μπορούν να χρησιμοποιηθούν για τη διευκόλυνση της αποτελεσματικής αναζήτησης και φιλτραρίσματος εγγράφων. Με την κατηγοριοποίηση εγγράφων με βάση τις λέξεις-κλειδιά τους, καθίσταται δυνατή η γρήγορη ανάκτηση πιθανών σχεδόν διπλότυπων παρουσιών ή το φιλτράρισμα του περιττού περιεχομένου. Οι τεχνικές αναζήτησης και φιλτραρίσματος που βασίζονται σε λέξεις-κλειδιά είναι καθοριστικές σε συστήματα ανίχνευσης αντιγραφής κειμένου.

Μια εναλλακτική μέθοδος, είναι αυτή της συν-εμφάνισης λέξεων όπου εξάγονται λέξεις-κλειδιά από τα έγγραφα, αλλά δεν φροντίζει για τη σειρά των λέξεων των προτάσεων και τη σημασία της λέξης από τα συμφραζόμενα. Η λεξιλογική βάση δεδομένων περιέχει την προκαθορισμένη δομή δέντρου λέξεων που αντιπροσωπεύει τη λέξη, το νόημα και τη σχέση με άλλες λέξεις. Προσδιορίζει μόνο την καλύτερη αντιστοίχιση ζευγών αντί να προσδιορίζει το ακριβές νόημα και αυτή η μέθοδος λειτουργεί με βάση το σώμα του κειμένου [DS22].

3.2.3.1 Ιστορική ανασκόπηση: Εύρεση διπλοτύπων με χρήση λέξεων κλειδιών

Οι [MGB14] αναφέρθηκαν σε μια προσέγγιση για τον εντοπισμό σχεδόν διπλότυπων ιστοσελίδων κατά την ανίχνευση ιστού. Εντοπίζονται σχεδόν διπλότυπες ιστοσελίδες, ακολουθούμενες από αποθήκευση ιστοσελίδων που έχουν ανιχνευτεί σε αποθετήρια. Οι λέξεις-κλειδιά εξάγονται από ανιχνευμένες σελίδες και υπολογίζεται ο βαθμός ομοιότητας μεταξύ δύο σελίδων. Τα έγγραφα θεωρούνται σχεδόν διπλότυπα εάν η βαθμολογία ομοιότητάς τους ικανοποιεί μια τιμή κατωφλίου.

3.2.3.2 Επίλογος λέξεων-κλειδιών

Συνοπτικά, οι λέξεις-κλειδιά είναι σημαντικά στοιχεία για τον εντοπισμό διπλότυπου κειμένου, βοηθώντας στην αναπαράσταση περιεχομένου, στη διάκριση, στον εντοπισμό δομικών προτύπων, στην εξαγωγή χαρακτηριστικών και στις διαδικασίες αναζήτησης/φιλτραρίσματος. Λαμβάνοντας υπόψη την παρουσία, τις παραλλαγές και την ευθυγράμμιση των λέξεων-κλειδιών, μπορεί κανείς να εντοπίσει

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αποτελεσματικά σχεδόν διπλότυπα στιγμιότυπα και να αποκτήσει πληροφορίες για τις θεματικές ομοιότητες και διαφορές μεταξύ των εγγράφων κειμένου. Η ενσωμάτωση λέξεων-κλειδιών στην ανάλυση ενισχύει την ακρίβεια και την αποτελεσματικότητα των συστημάτων ανίχνευσης αντιγραφής κειμένου.

3.2.4 Μηχανική μάθηση

Οι μέθοδοι μηχανικής μάθησης διαδραματίζουν κεντρικό ρόλο στην ανίχνευση αντιγραφής κειμένου, προσφέροντας ισχυρές τεχνικές για την αυτόματη αναγνώριση σχεδόν διπλότυπων παρουσιών και τη διάκρισή τους από το πρωτότυπο ή μοναδικό περιεχόμενο. Αυτές οι μέθοδοι αξιοποιούν προσεγγίσεις που βασίζονται σε δεδομένα για την εκμάθηση μοτίβων, ομοιοτήτων και ανομοιοτήτων εντός των εγγράφων κειμένου.

Οι μέθοδοι μηχανικής μάθησης διευκολύνουν την εξαγωγή πληροφοριακών χαρακτηριστικών από δεδομένα κειμένου. Διάφορες τεχνικές εξαγωγής χαρακτηριστικών, όπως bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), word embeddings ή αναπαραστάσεις n-gram, χρησιμοποιούνται για τη μετατροπή εγγράφων σε αριθμητικές ή διανυσματικές αναπαραστάσεις. Αυτά τα χαρακτηριστικά καταγράφουν τα βασικά χαρακτηριστικά του κειμένου, επιτρέποντας στα μοντέλα μηχανικής μάθησης να μαθαίνουν μοτίβα και ομοιότητες.

Τα μοντέλα μηχανικής μάθησης εκπαιδεύονται χρησιμοποιώντας δεδομένα με ετικέτα, τα οποία αποτελούνται από ζεύγη σχεδόν διπλότυπων και μη διπλότυπων στοιχείων. Μέσω της εκπαίδευσης, τα μοντέλα μαθαίνουν να εντοπίζουν μοτίβα και να κάνουν προβλέψεις με βάση τα εξαγόμενα χαρακτηριστικά. Οι δημοφιλείς αλγόριθμοι μηχανικής μάθησης για την ανίχνευση διπλότυπων κειμένων περιλαμβάνουν δέντρα αποφάσεων, random forest, support vector machines (SVM), λογιστική παλινδρόμηση (logistic regression) και μοντέλα βαθιάς μάθησης, όπως συνελικτικά νευρωνικά δίκτυα (CNN) ή αναδρομικά νευρωνικά δίκτυα (RNN).

Οι μέθοδοι μηχανικής μάθησης μπορούν να γενικευθούν από τα δεδομένα εκπαίδευσης για να ανιχνεύσουν σχεδόν διπλότυπα σε άγνωστα ή νέα έγγραφα. Αφού εκπαιδευτούν, τα μοντέλα μπορούν να επεξεργάζονται και να αναλύουν δεδομένα κειμένου που δεν έχουν εμφανιστεί, αξιοποιώντας τα μαθησιακά μοτίβα και τις ομοιότητες για τον εντοπισμό σχεδόν διπλών περιπτώσεων. Αυτή η ικανότητα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

γενίκευσης επιτρέπει την επεκτάσιμη ανίχνευση διπλασιασμού κειμένου σε μεγάλα σύνολα δεδομένων.

Ακόμα, μπορούν να εκτιμήσουν την ομοιότητα μεταξύ ζευγών εγγράφων κειμένου. Τα εκπαιδευμένα μοντέλα μπορούν να ορίσουν βαθμολογίες ομοιότητας ή πιθανότητες που υποδεικνύουν την πιθανότητα σχεδόν διπλασιασμού. Αυτές οι βαθμολογίες επιτρέπουν την κατάταξη ή ομαδοποίηση εγγράφων με βάση τα επίπεδα ομοιότητάς τους, βοηθώντας στον εντοπισμό και την οργάνωση σχεδόν διπλότυπων περιπτώσεων.

Άλλη μία παραλλαγή, είναι οι τεχνικές embedding που δημιουργούν διάνυσμα σταθερού μήκους και χρησιμοποιούνται για την κωδικοποίηση και την αναπαράσταση μιας οντότητας. Πειραματικά αποτελέσματα [MPK20] έδειξαν ότι οι τεχνικές μέτρησης ομοιότητας που χρησιμοποιούν μοντέλο μηχανικής μάθησης embedding για την αναπαράσταση του κειμένου σε αριθμητικά διανύσματα, καταγράφουν μεγαλύτερη ομοιότητα ενώ συγκρίνονται δύο παρόμοια τμήματα κειμένου, καθώς λαμβάνει επίσης υπόψη το σημασιολογικό νόημα.

Οι μέθοδοι συνόλου συνδυάζουν πολλαπλά μοντέλα μηχανικής μάθησης ή αλγόριθμους για να βελτιώσουν την ακρίβεια ανίχνευσης. Συγκεντρώνοντας τις προβλέψεις από πολλά μοντέλα, οι τεχνικές συνόλου μπορούν να επιτύχουν καλύτερη απόδοση και να χειριστούν διάφορους τύπους σχεδόν διπλότυπων περιπτώσεων.

3.2.4.1 Επίλογος μηχανικής μάθησης

Συνοπτικά, οι μέθοδοι μηχανικής μάθησης προσφέρουν ισχυρά εργαλεία για την ανίχνευση διπλότυπου κειμένου, διευκολύνοντας την εξαγωγή χαρακτηριστικών, εκπαίδευση μοντέλων, εκτίμηση ομοιότητας, γενίκευση σε άγνωστα δεδομένα και τεχνικές συνόλου. Αυτές οι μέθοδοι επιτρέπουν τον αυτοματοποιημένο εντοπισμό σχεδόν διπλότυπων περιπτώσεων, παρέχοντας επεκτάσιμες λύσεις για τη διατήρηση της ακεραιότητας και της ποιότητας του περιεχομένου κειμένου. Η αξιοποίηση των μεθόδων μηχανικής μάθησης ενισχύει την ακρίβεια, την αποτελεσματικότητα και την προσαρμοστικότητα των συστημάτων ανίχνευσης αντιγραφής κειμένου ενόψει του συνεχώς αυξανόμενου όγκου ψηφιακών πληροφοριών.

3.3 Επίλογος

Οι αλγόριθμοι για την ανίχνευση παρόμοιων εγγράφων είναι κρίσιμοι σε εφαρμογές όπου τα δεδομένα λαμβάνονται από πολλαπλές πηγές. Η αφαίρεση

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

παρόμοιων εγγράφων είναι απαραίτητη, όχι μόνο για τη μείωση του χρόνου εκτέλεσης, αλλά και για τη βελτίωση της ακρίβειας αναζήτησης. Σήμερα, οι ανιχνευτές μηχανών αναζήτησης ανακτούν δισεκατομμύρια μοναδικές διευθύνσεις URL, εκ των οποίων εκατοντάδες εκατομμύρια είναι διπλότυπα κάποιας μορφής [CFGM02].

Σε όλες τις τεχνικές που αναφέρθηκαν παραπάνω, όλα τα ζεύγη εγγράφων συγκρίνονται, δηλαδή, κάθε έγγραφο συγκρίνεται με κάθε άλλο έγγραφο και υπολογίζεται ένα βάρος ομοιότητας. Όμως, ο πραγματικός χρόνος εκτέλεσης εξαρτάται από τα δεδομένα και είναι δύσκολο να προβλεφθεί με ακρίβεια [CFGM02]. Αυτό έχει ως αποτέλεσμα, να μπορεί να αποβεί χρονοβόρα η ανίχνευση διπλοτύπων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

4 Μέθοδοι εντοπισμού ειδησεογραφικών άρθρων που αναφέρονται στην εξέλιξη στο χρόνο γνωστών γεγονότων

4.1 Εισαγωγή

Καθημερινά υπάρχει ένας τεράστιος όγκος νέων πληροφοριών στη διάθεσή μας, και ένα μεγάλο μέρος τους είναι ειδήσεις. Προέρχεται από πολλές διαφορετικές πηγές, συμπεριλαμβανομένων των παραδοσιακών μέσων όπως η εφημερίδα, το ραδιόφωνο και η τηλεόραση, και σύγχρονες πηγές όπως ο Ιστός. Χωρίς σωστή οργάνωση των συντριπτικών πληροφοριών, μπορεί κανείς εύκολα να χαθεί λόγω του τεράστιου μεγέθους του. Αυτό το φαινόμενο ονομάζεται υπερφόρτωση πληροφοριών. Οι άνθρωποι δεν αντιμετωπίζουν τα ειδησεογραφικά γεγονότα ως μεμονωμένα γεγονότα. Αντίθετα, τείνουν να συγκρίνουν νέες πληροφορίες με τη μνήμη και να τις εισάγουν στο υπάρχον δίκτυο δεδομένων, σε μια τοποθεσία δίπλα στις ήδη υπάρχουσες σχετικές ειδήσεις [FA07].

Δεν είναι εφικτό για έναν χρήστη να διαβάσει όλες τις πληροφορίες χωρίς καμία προεπεξεργασία, επειδή οι ειδήσεις που μπορεί να διαβάσει ένα άτομο είναι πολύ μικρότερες από την ποσότητα που παρήχθη στην ίδια χρονική περίοδο. Για να βοηθηθεί ο χρήστης να αποκτήσει τις απαραίτητες πληροφορίες στο συντομότερο χρονικό διάστημα, θα πρέπει να σχεδιαστεί ένα σύστημα που επεξεργάζεται αυτόματα τις ειδήσεις και τις μετατρέπει σε μια πιο αποδοτική μορφή για τον χρήστη.

Τα follow-up άρθρων ειδήσεων αναφέρονται σε επόμενα άρθρα ή ενημερώσεις που παρέχουν πρόσθετες πληροφορίες, αναλύσεις ή εξελίξεις που σχετίζονται με μια είδηση που δημοσιεύτηκε στο παρελθόν. Αυτές αποτελούν σημαντικό μέρος της δημοσιογραφίας, καθώς επιτρέπουν στους δημοσιογράφους να παρέχουν συνεχή κάλυψη και να ενημερώνουν το κοινό για τις τελευταίες εξελίξεις.

Χρειάζονται εξελίξεις γιατί μια ιστορία από μόνη της μπορεί να μην καλύπτει όλες τις πτυχές ενός γεγονότος σωστά. Οι δημοσιογράφοι δεν μπορούν να τα αναφέρουν όλα, πρέπει να επικεντρωθούν σε κομμάτια της ζωής και να τα αναφέρουν στους αναγνώστες τους σε ιστορίες που πρέπει να έχουν έκταση μισής σελίδας. Οι δημοσιογράφοι επιβάλλουν περιορισμούς χώρου και χρόνου στις αναφορές τους που δεν αντικατοπτρίζουν πάντα πόσο σημαντικό είναι το γεγονός στον πραγματικό κόσμο.

Οι δημοσιογράφοι προσπαθούν επίσης, να δείξουν τα follow-up γεγονότα σε αυτοτελή "κομμάτια" που ονομάζονται ειδήσεις. Υπάρχουν συχνά παράπλευρα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ζητήματα των ειδήσεων που δεν έχουν αγγίξει ή μεταγενέστερα γεγονότα που θα πρέπει να αναφερθούν επιπλέον.

4.2 Ανάλυση

Τα περιστατικά δεν αντιμετωπίζονται ως μεμονωμένες οντότητες, επειδή υπάρχουν ορισμένοι τύποι δεσμών μεταξύ των γεγονότων του πραγματικού κόσμου και οι άνθρωποι έχουν την ικανότητα να τα αναγνωρίσουν. Ένα περιστατικό είναι ένα συμβάν ειδήσεων που συμβαίνει σε μια συγκεκριμένη χρονική στιγμή ή σε ένα χρονικό εύρος, σε μια δεδομένη γεωγραφική τοποθεσία και περιλαμβάνει μία ή περισσότερες οντότητες και κάποια ενέργεια. Ένα περιστατικό μπορεί να εμφανιστεί σε διάφορα ρεπορτάζ ειδήσεων, και ακόμα κι αν τα λεξιλόγια δεν μοιάζουν, είναι πάντα το ίδιο περιστατικό αν μιλούν για το ίδιο πράγμα. Οι ειδήσεις δεν είναι στατικές, καθώς συχνά υπάρχουν νέες ενημερώσεις για ένα συγκεκριμένο θέμα. Οι ενημερώσεις δεν ανήκουν στο παλιό περιστατικό, αφού ο χρόνος ή/και άλλοι παράγοντες έχουν αλλάξει, αλλά υπάρχει κάποια εσωτερική σύνδεση που τις ενώνει [FA07].

Πρέπει να διακρίνουμε τις επόμενες ειδήσεις από αυτό που ονομάζουμε έκτακτες ειδήσεις, οι οποίες είναι αναφορές γεγονότων που εξακολουθούν να συμβαίνουν καθώς τα αναφέρουμε. Για παράδειγμα, οι ωριαίες αναφορές για μια αεροπειρατεία είναι μέρος μιας ιστορίας, η αναφορά της τελικής δίκης των αεροπειρατών είναι μια συνέχεια.

Επειδή τα γεγονότα συνδέονται συχνά, δεν είναι πάντα εύκολο να γνωρίζουμε τη διαφορά μεταξύ μιας συνέχειας και μιας νέας είδησης ή μιας ενημέρωσης μιας έκτακτης είδησης. Ωστόσο, ένα ιδιαίτερο χαρακτηριστικό μιας συνέχειας είναι ότι βασίζεται για τη σημασία ή το ενδιαφέρον της σε τουλάχιστον μία προηγούμενη είδηση. Μόνο και μόνο επειδή το follow-up άρθρο περιγράφει το αποτέλεσμα μιας προηγούμενης είδησης, δεν σημαίνει πως όλοι οι αναγνώστες θυμούνται την αρχική ιστορία, ακόμα κι αν την είδαν ή την άκουσαν [IH08].

4.2.1 Χρησιμότητα follow-up ειδήσεων

Τα follow-up δείχνουν πώς συνδέονται διάφορα μέρη της ζωής. Κάθε φορά που δημοσιεύεται ένα άρθρο, σε εκείνο το σημείο περιορίζεται την αναφορά του για το συμβάν σε μια μόνο χρονική στιγμή. Τα follow-up κάνουν την είδηση να αναφέρεται σε μεγαλύτερο χρονικό διάστημα και εξηγούν την αιτία και το αποτέλεσμα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Μια είδηση μπορεί να διεγείρει το ενδιαφέρον ή τη περιέργεια του αναγνώστη. Με θέματα ή γεγονότα που είναι αυτοτελή, μια καλογραμμένη αναφορά ειδήσεων ή ένα χαρακτηριστικό άρθρο αναφέρει όλα τα αναγκαία και σημαντικά γεγονότα. Ωστόσο, πολύ λίγα συμβάντα και θέματα μπορούν να καλυφθούν ολοτελώς με μία μόνο είδηση.

Λόγω έλλειψης χρόνου ή επειδή οι πηγές δεν ήταν διαθέσιμες, συχνά οι δημοσιογράφοι αναγκάζονται να υποβάλλουν ειδήσεις για δημοσίευση που δεν είναι σωστά ισορροπημένες. Τα follow-up δίνουν μια δεύτερη ευκαιρία να παραχθεί αυτή η ισορροπία. Η συνέχεια χρειάζεται να έχει τις πληροφορίες από την αρχική είδηση, αλλά θα πρέπει να έχει μια νέα οπτική γωνία ειδήσεων [IH08].

4.2.2 Σύνδεση ψηφιακών ειδήσεων στο Αρχείο Ψηφιακών Ειδήσεων

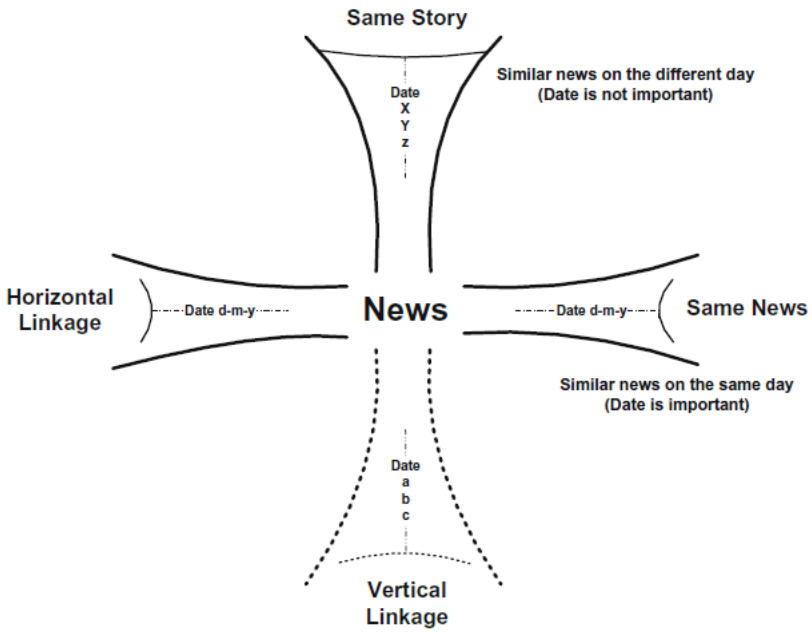
Το Αρχείο Ψηφιακών Ειδήσεων (Digital News Stories Archive DNSA) που δημιούργησαν οι [KRAA16], μέχρι και το 2018 διατηρούσε περισσότερες από χίλιες ειδήσεις μετά την διαγραφή διπλών διευθύνσεων URL και ειδήσεων, σε κάθε εξαγωγή. Το DNSA έχει άρθρα ειδήσεων από πολλές πηγές, και ήταν αναγκαίο να δημιουργηθεί ένας μηχανισμός που βοηθά τον αναγνώστη να διαβάσει ένα σύνολο σχετικών ειδήσεων για ένα γεγονός ή ζήτημα. Αυτή η σύνδεση οδηγεί τον αναγνώστη στην εύκολη περιήγηση μέσα από την τεράστια συλλογή. Χωρίς κατάλληλο και αποτελεσματικό μηχανισμό σύνδεσης για σχετικές ειδήσεις ή follow-up, το αρχείο εφημερίδων ή διαδικτυακών εφημερίδων δεν είναι τίποτα περισσότερο από μια συλλογή δεδομένων.

Σύμφωνα με τη μελέτη των [KUD18], ένας σύνδεσμος μπορεί να δημιουργηθεί με δύο τρόπους, δηλαδή κάθετη σύνδεση και οριζόντια σύνδεση.

- Κάθετη σύνδεση: Ο σύνδεσμος δημιουργείται μεταξύ των ειδήσεων που αναπαριστούν την «ίδια ιστορία» ανεξάρτητα από το χρονικό πλαίσιο με βάση όλες τις ειδήσεις που διατηρούνται στο DNSA. Η ίδια ιστορία μπορεί να ερμηνευθεί ως η επόμενη είδηση σχετικά με ένα ζήτημα από την ίδια πηγή ή από διαφορετικές πηγές.
- Οριζόντια σύνδεση: Ο σύνδεσμος δημιουργείται μεταξύ των ειδήσεων που αντιπροσωπεύουν τις «ίδιες ειδήσεις» με βάση όλες τις ειδήσεις που διατηρούνται στο DNSA σε μια συγκεκριμένη χρονική περίοδο, δηλαδή ημέρα, εβδομάδα, μήνα. Τα ίδια νέα μπορούν να ερμηνευτούν, όπως

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αναφέρθηκε προηγουμένως, ως επόμενες ειδήσεις για ένα θέμα από την ίδια πηγή ή από διαφορετικές πηγές.



Εικόνα 4-1: Η σύνδεση των ειδήσεων στο DNSA

Για να συνδεθούν άρθρα ψηφιακών ειδήσεων στο DNSA με βάση την ομοιότητα μεταξύ των ειδησεογραφικών άρθρων, οι επώνυμες οντότητες παίζουν πολύ σημαντικό ρόλο. Βοηθούν στον εντοπισμό σχετικών ειδήσεων από το αρχείο, με επεξεργασία των όρων που παρουσιάζονται στα άρθρα ειδήσεων. Τα ουσιαστικά, τα ρήματα, τα επιρρήματα κ.λπ. είναι διαφορετικοί όροι που σε συνδυασμό παρουσιάζουν το περιεχόμενο των ειδήσεων. Στα ακαδημαϊκά περιοδικά, τα ουσιαστικά θεωρούνται οι κύριες φράσεις-κλειδιά αλλά και άλλοι όροι όπως ρήματα και επιρρήματα παίζουν επίσης ζωτικό ρόλο στην αναπαράσταση των ειδήσεων. Βασιζόμενοι στα παραπάνω, ο “Similarity-based on Named Entities in Stories (SNES)” αλγόριθμος προτάθηκε από τους [KUUN20], κάνοντας χρήση επώνυμων οντοτήτων που αναφέρονται στα άρθρα ειδήσεων για τη σύνδεση follow-up ειδήσεων.

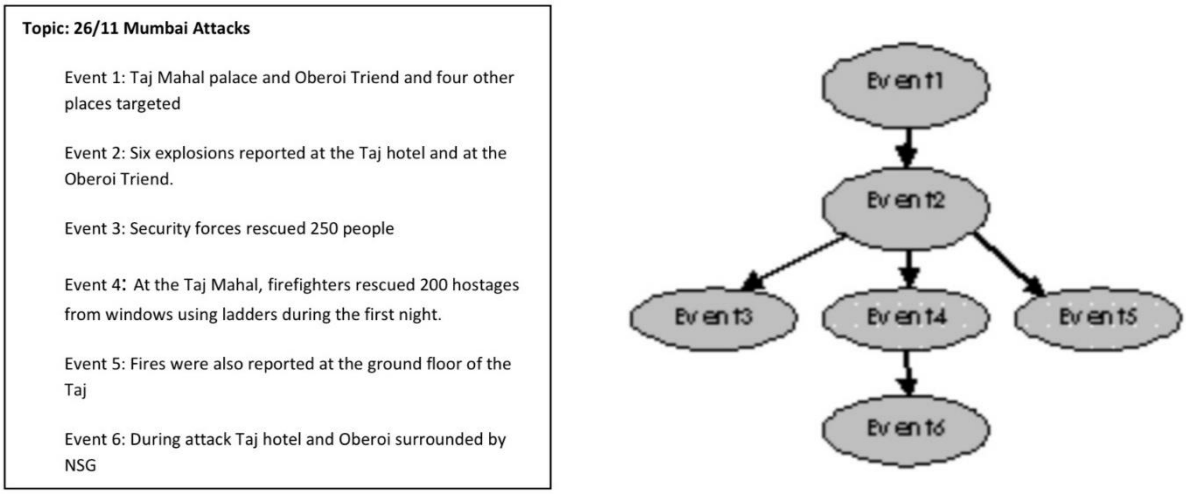
4.2.3 Ανίχνευση και παρακολούθηση θεμάτων (Topic detection and tracking (TDT))

Η ανάκτηση ειδήσεων για το ίδιο θέμα από πολλές πηγές και η ενημέρωση των πληροφοριών χρησιμοποιούν τεχνικές που είναι ικανές να εξάγουν τη δομή των ειδήσεων. Βοηθούν στην κατανόηση της εξέλιξης των γεγονότων για το ίδιο θέμα. Υπάρχουν διαφορετικές τεχνικές ανάκτησης πληροφοριών, όπως η ανίχνευση και παρακολούθηση θεμάτων (TDT), η παρακολούθηση συμβάντων και οι προσεγγίσεις

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

εξέλιξης συμβάντων. Το TDT είναι ένας ενεργός ερευνητικός τομέας τα τελευταία χρόνια. Το TDT έχει διαφορετικές τεχνικές για τον εντοπισμό θεμάτων ειδήσεων και την παρακολούθηση ειδήσεων για ένα θέμα ειδήσεων [FA09].

Οι τεχνικές TDT προσπαθούν να ανιχνεύσουν ή να ομαδοποιήσουν ειδήσεις σε γεγονότα, χωρίς να ορίσουν ή να ερμηνεύσουν τη συσχέτιση μεταξύ αυτών των γεγονότων. Επίσης, η εξέλιξη γεγονότων είναι μια νέα ιδέα που αναπτύχθηκε τη προηγούμενη δεκαετία και είναι η διαδικασία ανάπτυξης μετάβασης σχετικών γεγονότων εντός του ίδιου θέματος. Ένα γράφημα εξέλιξης γεγονότων κατασκευάζεται για να παρουσιάσει τη σχέση μεταξύ γεγονότων. Απαιτούνται συμβάντα για την κατασκευή γραφήματος εξέλιξης γεγονότων [F08]. Η παρακολούθηση γεγονότων γίνεται χρησιμοποιώντας το μοτίβο εξέλιξης γεγονότων γνωστό ως EPET (event pattern-supported event tracking).



Εικόνα 4-2: Παράδειγμα γραφήματος εξέλιξης γεγονότων. Πηγή: [NHD12]

Το παραπάνω παράδειγμα, σχετίζεται με το θέμα ειδήσεων "26/11 Mumbai Attacks" ως ενδεικτικό. Είναι μια λίστα με 6 ειδησεογραφικά συμβάντα σε αυτό το θέμα που παρακολουθούνται από την τεχνική TDT χωρίς να παρουσιάζεται πώς αυτά τα συμβάντα εξελίσσονται ως γράφημα εξέλιξης γεγονότων. Με περιήγηση στη λίστα γίνεται κατανοητό τι συνέβη σύμφωνα με τη χρονική σειρά. Τα συμβάντα της λίστας αναπαρίστανται σε γραφική μορφή χρησιμοποιώντας το γράφημα εξέλιξης συμβάντων [NHD12].

Το TDT ασχολείται με τον εντοπισμό και την παρακολούθηση γεγονότων. Η συμβολή σε αυτή τη διαδικασία είναι μια ροή ιστοριών. Αυτή η ροή μπορεί να είναι ή

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

να μην είναι εκ των προτέρων τμηματοποιημένη σε ιστορίες και τα γεγονότα μπορεί να είναι ή να μην είναι γνωστά στο σύστημα. Αυτό οδηγεί στον ορισμό τριών τεχνικών που πρέπει να αντιμετωπιστούν στη μελέτη TDT. Αυτά είναι συγκεκριμένα η παρακολούθηση γνωστών γεγονότων, ο εντοπισμός άγνωστων γεγονότων και η τμηματοποίηση μιας είδησης.

Οι εργασίες στο TDT περιλαμβάνουν την τμηματοποίηση ιστορίας, τον εντοπισμό θεμάτων, τον εντοπισμό νέων συμβάντων, τον εντοπισμό συνδέσμων και την παρακολούθηση θεμάτων.

- Ο στόχος της τμηματοποίησης της ιστορίας είναι να καθοριστούν τα όρια για συνεκτικά τμήματα κειμένου.
- Η ανίχνευση θεμάτων ομαδοποιεί ιστορίες του ίδιου θέματος.
- Η ανίχνευση νέων συμβάντων λαμβάνει δυαδική απόφαση για κάθε νέο έγγραφο εάν συζητά ένα νέο θέμα που δεν έχει ακόμη αναφερθεί.
- Η ανίχνευση συνδέσμου καθορίζει εάν δύο έγγραφα είναι του ίδιου θέματος.
- Η παρακολούθηση θεμάτων παρακολουθεί το θέμα για να βρει όλες τις ακόλουθες ιστορίες.

Το TDT γενικά εστιάζει στην ανίχνευση θεμάτων και νέων γεγονότων καθώς και ειδήσεων σε διαφορετικά θέματα. Η αλληλεπίδραση μεταξύ των γεγονότων είναι εκτός του πεδίου εφαρμογής της TDT έρευνας, δηλαδή η εξέλιξη γεγονότων απαιτεί περαιτέρω έρευνα. Τα γραφήματα εξέλιξης γεγονότων δείχνουν πως τα περίπλοκα συμβάντα αλληλοεπιδρούν στη γραφική δομή για εύκολη πλοήγηση και περιήγηση [NHD12].

4.2.4 Ιστορική ανασκόπηση: προσεγγίσεις εύρεσης follow up

Οι [CSS02], [APL98], [YPC98] χρησιμοποίησαν τη προσέγγιση επεξεργασίας φυσικής γλώσσας με συνδυασμό λεξιλογικών αλυσίδων με λέξεις-κλειδιά για παρακολούθηση γεγονότων και εξήγαγαν επτά τύπους επώνυμων οντοτήτων. Ερευνούν μια νέα προσέγγιση που αποτελείται από τα ακόλουθα στοιχεία:

1. Ταξινόμηση εγγράφων σε γενικά θέματα καθένα από τα οποία αποτελείται από πολλαπλά γεγονότα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2. Προσδιορισμός επώνυμων οντοτήτων, βελτιστοποίηση του βάρους σε σχέση με τις κανονικές λέξεις για κάθε θέμα και υπολογισμός λίστας stop words ανά γεγονός.
3. Για ομοιότητα περιεχομένου μεταξύ δύο γεγονότων είναι απαραίτητη η χρονική εγγύτητα και εγγύτητα εγγράφων.

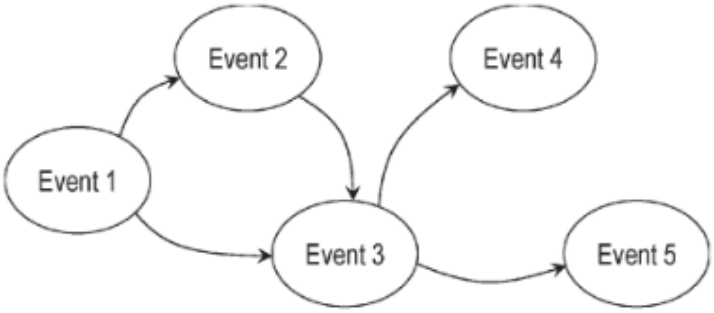
Η εργασία [M03] ήταν η πρώτη έρευνα για εξέλιξη γεγονότος. Τα άρθρα ειδήσεων που αναφέρονται σε ένα θέμα είναι χρονικά γραμμικά διατεταγμένα. Ένα σημαντικό γεγονός μπορεί να οδηγήσει σε πολλά άλλα γεγονότα. Τα γεγονότα στην αρχή μπορεί να έχουν περισσότερη επιρροή στα γεγονότα που έρχονται έπειτα. Λαμβάνοντας υπόψη ένα γεγονός στη χρονική διάταξη, γίνεται αντιληπτή η εξέλιξη των γεγονότων που αναφέρονται σε ένα περιστατικό. Τα γεγονότα και η εξέλιξη των γεγονότων μπορούν να αναπαρασταθούν ως δομή γραφήματος. Χρησιμοποιήθηκαν επώνυμες οντότητες, συμπεριλαμβανομένων τοποθεσιών, ονομάτων και χρονικών στιγμών, για μέτρηση της ομοιότητας των γεγονότων. Υπάρχει έλλειψη λεπτομερειών στο μοντέλο εξέλιξης γεγονότων και αποτυχία παρουσίασης πειραματικών αποτελεσμάτων αξιολόγησης της μεθόδου.

Η εργασία [NFPA04] καθόρισε την έννοια του threading γεγονότων, λαμβάνοντας υπόψη ένα μικρό αριθμό εγγράφων και συμβάντων σε ένα θέμα ειδήσεων. Ο δικός τους ορισμός του threading συμβάντων είναι κοντά στην εξέλιξη του συμβάντος εκτός από το ότι θεωρούν το threading συμβάντων ως δομή δέντρου παρά ένα γράφημα. Η έννοια του threading συμβάντων είναι πράγματι μόνο μέρος της έννοιας της εξέλιξης των γεγονότων, και η προτεινόμενη τεχνική τους δεν μπορεί να αντιμετωπίσει την πολυπλοκότητα σε σχέσεις εξέλιξης γεγονότων. Οι μέθοδοι της προτεινόμενης τεχνικής λαμβάνουν υπόψη μόνο την ομοιότητα μεταξύ γεγονότων. Τέτοιες μέθοδοι δεν είναι αποτελεσματικές και επαρκείς για τον προσδιορισμό της εξέλιξης του γεγονότος.

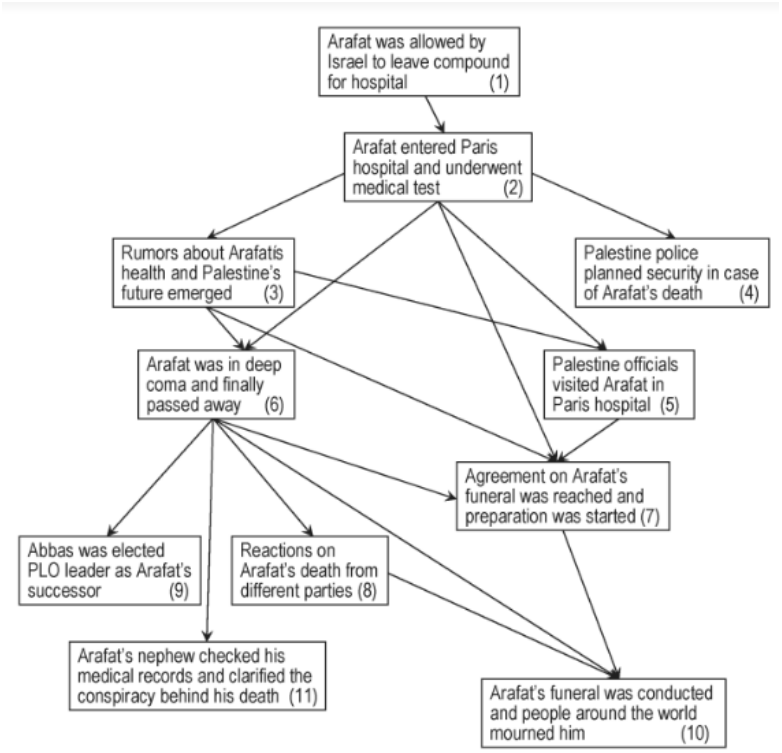
Η εργασία [WC07] πρότεινε μια τεχνική ανακάλυψης προτύπων εξέλιξης γεγονότος που προσδιορίζει συμβάντα μαζί με τις χρονικές τους σχέσεις που συμβαίνουν συχνά σε μια συλλογή γεγονότων του ίδιου τύπου. Η συγκεκριμένη μελέτη εστιάζει στην κατάτμηση μιας ακολουθίας ειδήσεων του συγκεκριμένου γεγονότος και τη σχέση τους. Ωστόσο, οι σχέσεις μεταξύ των συμβάντων αποτελούνται μόνο από χρονικές στιγμές, που δίνουν κατά προσέγγιση αποτελέσματα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Η εργασία [YSW09] πρότεινε μία τεχνική ανακάλυψης μοτίβων εξέλιξης γεγονότων που προσδιορίζει συμβάντα με τις χρονικές τους σχέσεις που συμβαίνουν συχνά σε συλλογή γεγονότων του ίδιου τύπου. Η μελέτη διαφέρει από προηγούμενες μελέτες στο ότι εστιάζει στην κατάτμηση ακολουθίας ειδήσεων ενός συγκεκριμένου γεγονότος σε περιστατικά και γενίκευση περιστατικών σε διάφορα παρόμοια θέματα. Ωστόσο, οι σχέσεις μεταξύ των συμβάντων αποτελούνται μόνο από χρονικές στιγμές. Σε αυτή τη μελέτη ορίζεται επίσημα η εξέλιξη του γεγονότος. Επιπλέον, εισάγεται η χρονική σχέση, ομοιότητα συμβάντος, και χρονική εγγύτητα για τον προσδιορισμό της εξέλιξης του γεγονότος για την κατασκευή των γραφημάτων εξέλιξης.



Εικόνα 4-3: Παράδειγμα γραφήματος εξέλιξης γεγονότων. Πηγή: [YSW09]



Εικόνα 4-4: Γράφημα εξέλιξης γεγονότων για το θέμα «Ο θάνατος του Αραφάτ». Οι αριθμοί στις αγκύλες υποδεικνύουν τις χρονικές ταξινομήσεις των γεγονότων που αντιπροσωπεύονται από τα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ορθογώνια. Οι άκρες μεταξύ δύο ορθογωνίων δείχνουν την εξέλιξη του γεγονότος και τη σχέση μεταξύ των υποκείμενων γεγονότων τους.

4.3 Επίλογος

Λειτουργώντας ως αγωγός που συνδέει το παρελθόν με το παρόν, οι επακόλουθες αφηγήσεις ειδήσεων έχουν τεράστια αξία στην κατανόησή μας περίπλοκων κοινωνικοπολιτικών φαινομένων. Εν μέσω της αδιάκοπης εισροής πληροφοριών, οι follow-up ειδήσεις χρησιμεύουν ως ζωτικός μηχανισμός για τη διατήρηση και την εμβάθυνση της κατανόησης σημαντικών γεγονότων και ζητημάτων.

Επιπλέον, παρέχουν μια ολοκληρωμένη κατανόηση περίπλοκων ζητημάτων που συχνά αποφεύγουν την επιφανειακή κατανόηση. Με την επανεξέταση προηγούμενων αφηγήσεων, καθίσταται δυνατό να διακρίνουμε τις μόνιμες προεκτάσεις, τις μακροπρόθεσμες συνέπειες και τις συνεχείς εξελίξεις που προκύπτουν από σημαντικά γεγονότα ή αποφάσεις.

Συνοψίζοντας, οι follow-up ειδήσεις προστατεύουν από την απώλεια κρίσιμων αφηγήσεων, δημιουργεί ολοκληρωμένη κατανόηση περίπλοκων ζητημάτων και δίνει τη δυνατότητα στα άτομα να διαμορφώνουν ενεργά το κοινωνικό τους περιβάλλον.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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

5 Ανάλυση και σχεδίαση του νέου συστήματος

5.1 Εισαγωγή

Οι βάσεις δεδομένων ειδήσεων παρέχουν ένα οργανωμένο και προσβάσιμο αποθετήριο πληροφοριών χρήσιμο για ένα ευρύ φάσμα σκοπών, από την ιστορική ανάλυση έως την ειδησεογραφία και την έρευνα σε πραγματικό χρόνο. Παρέχουν στους χρήστες τα μέσα για την άντληση πληροφοριών, την επαλήθευση πληροφοριών και τη συμβολή στη λήψη τεκμηριωμένων αποφάσεων σε διάφορους τομείς.

5.2 Συνοπτική περιγραφή του πληροφοριακού συστήματος

Στο δυναμικό πλαίσιο των μέσων μαζικής ενημέρωσης και της δημοσιογραφίας, η δημιουργία μιας ισχυρής βάσης δεδομένων ειδήσεων απαιτεί μια σε βάθος προσέγγιση πολλαπλών σταδίων. Η διαδικασία αυτή διασφαλίζει ότι οι ειδήσεις συλλέγονται, οργανώνονται και βελτιώνονται συνεχώς, καθιστώντας την πολύτιμη πηγή για τους ερευνητές, τους δημοσιογράφους και τους αναλυτές. Η διαδικασία δημιουργίας της βάσης δεδομένων ειδήσεων και γεγονότων στην παρούσα εργασία συνοψίζεται συνοπτικά παρακάτω:

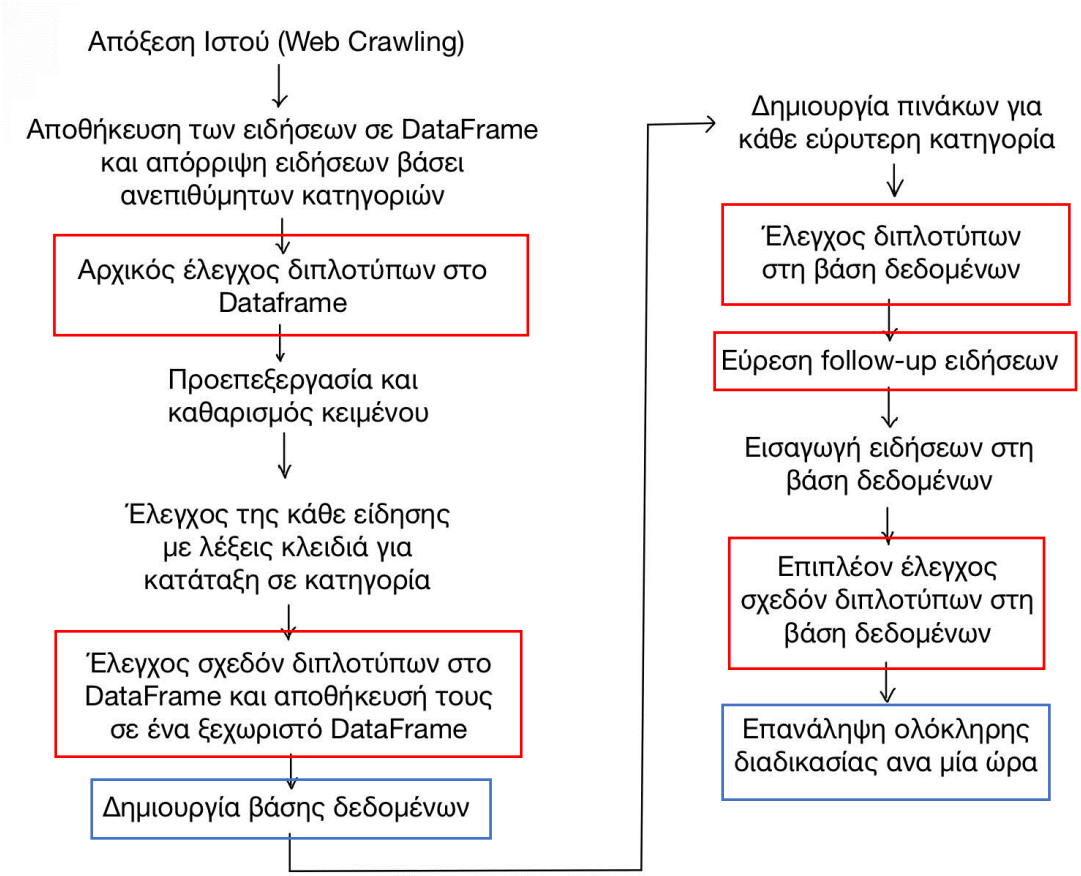
1. **Απόξεση Ιστού (Web Crawling)**
2. **Αποθήκευση των ειδήσεων σε DataFrame και απόρριψη ειδήσεων βάσει ανεπιθύμητων κατηγοριών**
3. **Αρχικός έλεγχος διπλότυπων στο DataFrame**: Οι δημιουργούμενες συναρτήσεις εντοπίζουν διπλότυπα ειδήσεων εντός του DataFrame, διασφαλίζοντας την ακρίβεια των πληροφοριών και ελαχιστοποιώντας τον πλεονασμό.
4. **Προεπεξεργασία και καθαρισμός κειμένου**
5. **Έλεγχος της κάθε είδησης με λέξεις-κλειδιά για κατάταξη σε κατηγορία**
6. **Έλεγχος σχεδόν διπλότυπων στο DataFrame και αποθήκευση τους σε ένα ξεχωριστό DataFrame:** Οι δημιουργούμενες συναρτήσεις εντοπίζουν σχεδόν διπλότυπα ειδήσεων εντός του DataFrame, και τα διαχωρίζουν σε ένα εξειδικευμένο Dataframe που σχετίζεται με την αποθήκευση διπλοτύπων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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**
9. **Έλεγχος διπλότυπων στη βάση δεδομένων**: Αυστηροί έλεγχοι εντός της βάσης δεδομένων εντοπίζουν και εξαλείφουν τυχόν διπλότυπα, διασφαλίζοντας την ακεραιότητα των δεδομένων.
10. **Εύρεση follow-up ειδήσεων**: Κατάλληλες συναρτήσεις εντοπίζουν τις follow-up ειδήσεις, συνδέοντας τις εξελισσόμενες αφηγήσεις και παρέχοντας ολοκληρωμένη εικόνα των γεγονότων.
11. **Εισαγωγή ειδήσεων στη βάση δεδομένων**
12. **Επιπλέον έλεγχος σχεδόν διπλότυπων στη βάση δεδομένων**: Παρόμοια με τα προηγούμενα στάδια, διεξάγεται ένας τελικός έλεγχος για σχεδόν διπλότυπα εντός της βάσης δεδομένων για τη βελτίωση της συνολικής ποιότητάς της.
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.

Τα παραπάνω βήματα οπτικοποιούνται στο παρακάτω διάγραμμα για την καλύτερη κατανόηση της ροής της διαδικασίας. Με κόκκινο πλαίσιο αναγράφονται οι διαδικασίες που πραγματεύονται μόνο στην παρούσα εργασία - του προαναφερθέντος, κοινού μεγαλύτερου έργου που εκπονήσαμε με τη Σ. Σκορδίλη - ενώ με μπλε πλαίσια υποδηλώνονται διαδικασίες που απαραίτητα χρησιμοποιούνται και είναι άξιο να αναφερθούν στην παρούσα πτυχιακή εργασία, όπως εξετάζονται επίσης και στην εργασία [Σ23]. Τέλος, οι διαδικασίες που δεν έχουν πλαίσιο περιγράφονται μόνο στην εργασία [Σ23].

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

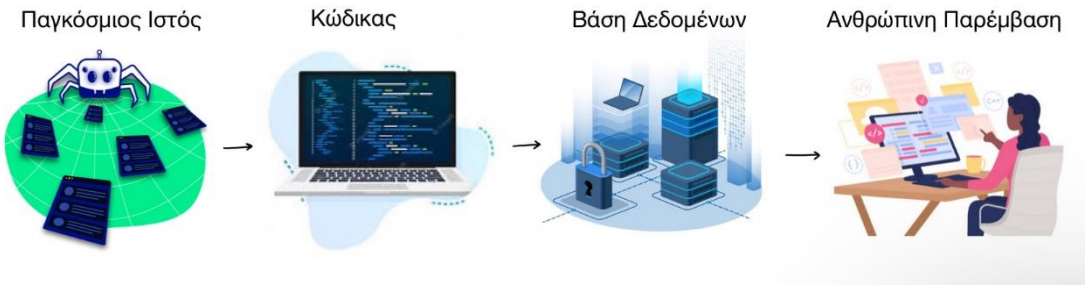


Γράφημα 5-1: Διάγραμμα σχεδίασης της ροής του προτεινόμενου πληροφοριακού συστήματος

5.3 Τρόπος λειτουργίας Πληροφοριακού Συστήματος

Ένα Πληροφοριακό Σύστημα είναι το θεμέλιο των οργανωτικών διαδικασιών, παρέχοντας το πλαίσιο μέσω του οποίου τα δεδομένα συλλέγονται, επεξεργάζονται, αποθηκεύονται και μετατρέπονται σε χρήσιμες πληροφορίες. Είναι μια δυναμική αλληλεπίδραση τεχνολογίας, ανθρώπων και διαδικασιών, που συνεργάζονται για να διευκολύνουν αποτελεσματικές λειτουργίες, ενημερωμένη λήψη αποφάσεων και στρατηγική ανάπτυξη.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Γράφημα 5-1: Τρόπος λειτουργίας Πληροφοριακού Συστήματος. Πηγές: www.istockphoto.com , www.freepik.com , https://www.scraping-bot.io/how-to-build-a-web-crawler/

Όπως αναγράφεται στην παραπάνω εικόνα, η διαδικασία ξεκινά από τη συλλογή από τον Παγκόσμιο Ιστό, συνεχίζει με το κομμάτι του κώδικα για κάθε είδος επεξεργασίας, οι πληροφορίες έπειτα μεταφέρονται σε βάση δεδομένων και τέλος λαμβάνει χώρα η ανθρώπινη παρέμβαση για να διασφαλιστεί η ακρίβεια των πληροφοριών. Η διαδικασία αυτή αντιπροσωπεύει την ημιαυτόματη ροή εργασίας ενός συστήματος πληροφοριών.

Ο Παγκόσμιος Ιστός αποτελεί το παγκόσμιο δίκτυο συνδεδεμένων εγγράφων και πόρων, τα οποία είναι προσβάσιμα μέσω του Διαδικτύου και διαδραματίζει το ρόλο μιας τεράστιας αποθήκης πληροφοριών που καλύπτουν ένα ευρύ φάσμα θεμάτων. Οι χρήστες του διαδικτύου διαθέτουν τη δυνατότητα να αναζητούν και να έχουν πρόσβαση σε ιστότοπους, ιστοσελίδες και περιεχόμενο με πολυμέσα, προσφέροντας έτσι ένα πλήθος πληροφοριών για τις ανάγκες της έρευνας και της ενημέρωσης.

Ο κώδικας αναφέρεται στη διαδικασία σύνταξης οδηγιών σε μια γλώσσα προγραμματισμού για τη δημιουργία εφαρμογών λογισμικού ή λειτουργιών. Στο πλαίσιο ενός πληροφοριακού συστήματος, οι προγραμματιστές δημιουργούν εφαρμογές ή εργαλεία που βασίζονται στον ιστό για τη συλλογή, επεξεργασία και παρουσίαση πληροφοριών από τον Ιστό στους χρήστες.

Οι βάσεις δεδομένων λειτουργούν ως δομημένοι χώροι αποθήκευσης και διαχείρισης πληροφοριών. Παρέχουν έναν συστηματικό και τακτοποιημένο τρόπο αποθήκευσης και ανάκτησης δεδομένων, διευκολύνοντας την αποτελεσματική διαχείριση πληροφοριών. Όταν πληροφορίες συλλέγονται από τον Παγκόσμιο Ιστό και υπόκεινται σε επεξεργασία μέσω προγραμματισμού, η αποτελεσματική τους αποθήκευση και ανάκτηση διευκολύνεται μέσω των βάσεων δεδομένων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Παρά την πολυπλοκότητα της κωδικοποίησης και των βάσεων δεδομένων, συχνά υπάρχουν πολυπλοκότητες ή ασάφειες στα δεδομένα που απαιτούν ανθρώπινη κρίση και παρέμβαση. Οι άνθρωποι διαδραματίζουν κρίσιμο ρόλο στη διασφάλιση της ακρίβειας, της συνάφειας και του πλαισίου των δεδομένων. Αυτό περιλαμβάνει καθήκοντα όπως επικύρωση δεδομένων, καθαρισμός, διασφάλιση ποιότητας, ακόμη και λήψη υποκειμενικών αποφάσεων κατά την ερμηνεία πληροφοριών που μπορεί να μην είναι πλήρως κατανοητές από αυτοματοποιημένες διαδικασίες. Η ανθρώπινη επαλήθευση προσθέτει ένα ουσιαστικό επίπεδο ποιοτικού ελέγχου, διασφαλίζοντας ότι οι πληροφορίες που παρουσιάζονται στους χρήστες είναι ακριβείς και ουσιαστικές.

Καταλήγοντας, μια συνδυαστική προσέγγιση που ενώνει την αποδοτικότητα του αυτοματισμού με την κριτική σκέψη και τη λογική των ανθρώπων μπορεί να δημιουργήσει μια ισχυρή και αξιόπιστη βάση δεδομένων για ειδησεογραφικά γεγονότα. Ο συνδυασμός αυτός, αξιοποιεί την τεχνολογία για τη διαχείριση μεγάλων όγκων εργασίας, ενώ ταυτόχρονα διασφαλίζει την ακρίβεια, την ποιότητα και την κατανόηση του συμφραζόμενου, τα οποία μόνο η ανθρώπινη παρέμβαση μπορεί να προσφέρει.

5.4 Ανάλυση modules, πακέτων, βιβλιοθηκών, drivers και frameworks του κώδικα σε γλώσσα προγραμματισμού Python

Προτού ξεκινήσει η ανάλυση της συγκεκριμένης ενότητας, απαραίτητη είναι η ανάλυση και επεξήγηση κάποιων εννοιών. Συγκεκριμένα, θα πρέπει να ξεκαθαριστούν οι όροι script, module, package, library, driver, framework, που χρησιμοποιούνται στον κώδικα της παρούσας εργασίας.

- ***Script:*** Ένα Python script αποτελεί ένα αυτόνομο αρχείο που περιλαμβάνει εντολές που εκτελούνται απευθείας. Σε ένα script, ο κώδικας εκτελείται προκειμένου να φέρει εις πέρας μια συγκεκριμένη λειτουργία. Συνήθως, τα scripts περιέχουν λογική που δεν είναι οργανωμένη σε κλάσεις ή συναρτήσεις.
- ***Module:*** Ένα Python module είναι ένα αρχείο που σχεδιάζεται για να εισαχθεί σε άλλα scripts ή modules. Ένα module συνήθως περιλαμβάνει στοιχεία όπως συναρτήσεις, μεταβλητές και κλάσεις που προορίζονται να χρησιμοποιηθούν από άλλα αρχεία που το εισάγουν με τη δήλωση `import`.
- ***Package:*** Ένα πακέτο στην Python είναι μια οργανωμένη συλλογή σχετικών modules που συνεργάζονται για να παρέχουν συγκεκριμένες λειτουργίες. Τα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

modules αυτά αποθηκεύονται σε έναν κατάλογο και μπορούν να εισαχθούν στον κώδικα όπως άλλα modules. Συχνά, ο φάκελος πακέτου περιλαμβάνει ένα ειδικό αρχείο `__init__`, το οποίο δηλώνει ότι πρόκειται για ένα πακέτο.

- ***Library:*** Στην Python, ο όρος "library" αναφέρεται συνήθως σε μια συλλογή από modules ή πακέτα που παρέχουν έτοιμο κώδικα για την εκτέλεση κοινών εργασιών. Οι βιβλιοθήκες μπορεί να περιλαμβάνουν δεκάδες ή ακόμη και εκατοντάδες μεμονωμένα modules που προσφέρουν διάφορες λειτουργίες.
- ***Driver:*** Ένας driver αποτελεί ένα σύνολο αρχείων που δίνουν οδηγίες σε ένα κομμάτι υλικού για τον τρόπο λειτουργίας του, επικοινωνώντας με το λειτουργικό σύστημα του υπολογιστή.
- ***Framework:*** Ένα framework στην Python αντιπροσωπεύει μια προκαθορισμένη δομή ή αρχιτεκτονική που λειτουργεί ως πλατφόρμα, επιτρέποντας την ανάπτυξη λογισμικού χωρίς την ανάγκη δημιουργίας κάθε λειτουργικότητας από την αρχή. Ένα framework παρέχει έναν τρόπο να προστεθούν και να προσαρμοστούν επιπλέον χαρακτηριστικά και λειτουργίες για τη δημιουργία σύνθετων έργων.

Για την εκτέλεση της υλοποίησης του κώδικα ήταν απαραίτητο μεγάλο πλήθος βιβλιοθηκών, πακέτων, modules, drivers και frameworks ανάλογα με τις επιθυμητές λειτουργίες της υλοποίησης. Βέβαια, η εγκατάστασή τους πρέπει να γίνει σε συγκεκριμένο φάκελο προκειμένου να εκτελεστούν σωστά. Αν στο τρέχοντα φάκελο (.) βρίσκεται το αρχείο με τον κώδικα, τότε οι βιβλιοθήκες, τα πακέτα, τα modules, οι drivers και τα frameworks θα πρέπει να εγκαθίστανται στο μονοπάτι *./venv/Scripts*.

Παρακάτω παρουσιάζονται τα εργαλεία που απαιτήθηκαν για το τμήμα του κώδικα που αναπτύχθηκε σε αυτήν την πτυχιακή εργασία, μαζί με τις σχετικές εντολές που χρησιμοποιήθηκαν για τη φόρτωσή τους.

5.4.1 Modules

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

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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

SciPy: είναι ένα λογισμικό ανοιχτού κώδικα για τα μαθηματικά, την επιστήμη και τη μηχανική. Περιλαμβάνει ενότητες για στατιστικές, βελτιστοποίηση (optimization), ολοκλήρωση (integration), γραμμική άλγεβρα, μετασχηματισμούς Fourier, επεξεργασία σήματος και εικόνας και άλλα.

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install scipy* [SciD23].

Levenshtein: είναι επέκταση της Python για υπολογισμό αποστάσεων και ομοιοτήτων επεξεργασίας συμβολοσειρών.

Για την φόρτωση του συγκεκριμένου module απαραίτητη είναι η χρήση της εντολής στο command line *pip install python-Levenshtein* [B23a].

Difflib: Η difflib στην Python παρέχει κλάσεις και συναρτήσεις για σύγκριση ακολουθιών, συγκεκριμένα, συμβολοσειρών. Η πιο συχνά χρησιμοποιούμενη κλάση στο difflib είναι η ***SequenceMatcher***. Παρέχει μεθόδους σύγκρισης ακολουθιών, όπως συμβολοσειρές, λίστες ή άλλα επαναλαμβανόμενα αντικείμενα. Η κλάση SequenceMatcher υπολογίζει την αναλογία ομοιότητας μεταξύ δύο ακολουθιών χρησιμοποιώντας τον αλγόριθμο Ratcliff/Obershelp, ο οποίος υπολογίζει μια βαθμολογία ομοιότητας.

Για τη φόρτωση του συγκεκριμένου module δεν απαιτείται ξεχωριστή εγκατάσταση χρησιμοποιώντας pip. Το difflib είναι μέρος της τυπικής βιβλιοθήκης Python, επομένως μπορεί να χρησιμοποιηθεί χωρίς πρόσθετα βήματα εγκατάστασης.

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 Packages

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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Στόχος του είναι 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: Το 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 Libraries

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].

Itertools: Η βιβλιοθήκη itertools της Python μπορεί να συνθέσει κομψές λύσεις για μια ποικιλία προβλημάτων με τις λειτουργίες που παρέχει. Με την more-itertools

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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* [R23].

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].

FuzzyWuzzy: είναι μια βιβλιοθήκη της Python που χρησιμοποιείται για αντιστοίχιση συμβολοσειρών. Η αντιστοίχιση συμβολοσειρών είναι η διαδικασία εύρεσης συμβολοσειρών που ταιριάζουν με ένα δεδομένο μοτίβο. Βασικά χρησιμοποιεί την απόσταση Levenshtein για να υπολογίσει τις διαφορές μεταξύ των ακολουθιών.

Για την φόρτωση της συγκεκριμένης βιβλιοθήκης απαραίτητη είναι η χρήση της εντολής στο command line *pip install fuzzywuzzy* [S22].

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].

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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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

5.4.5 Frameworks

Sentence Transformers: Αυτό το πλαίσιο παρέχει μια εύκολη μέθοδο υπολογισμού πυκνών διανυσματικών αναπαραστάσεων για προτάσεις, παραγράφους και εικόνες. Τα μοντέλα βασίζονται σε δίκτυα μετασχηματιστών, όπως το BERT, και επιτυγχάνουν state-of-the-art επιδόσεις σε διάφορες εργασίες. Το κείμενο ενσωματώνεται σε διανυσματικό χώρο έτσι ώστε παρόμοιο κείμενο να είναι κοντά και να μπορεί να βρεθεί αποτελεσματικά χρησιμοποιώντας το cosine similarity.

Για την φόρτωση του συγκεκριμένου framework απαραίτητη είναι η χρήση της εντολής στο command line *pip install -U sentence-transformers* [RM23][R22].

5.5 Επίλογος

Εν κατακλείδι, πραγματοποιήθηκε η σχεδίαση και η ανάλυση του πληροφοριακού συστήματος της συγκεκριμένης εργασίας. Με άλλα λόγια, επεξηγήθηκαν επιγραμματικά τα βήματα που ακολουθήθηκαν στον κώδικα, καθώς και ο τρόπος λειτουργίας του συστήματος αυτού, από τον Παγκόσμιο Ιστό έως το τελικό προϊόν. Επιπλέον, παρουσιάστηκαν οι βιβλιοθήκες, τα πακέτα, τα modules, οι drivers και τα frameworks που ήταν αναγκαία για την πραγματοποίηση των λειτουργιών του κώδικα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

6 Δημιουργία βάσης δεδομένων

6.1 Εισαγωγή

Οι βάσεις δεδομένων αποτελούν την κύρια δομή αποθήκευσης δεδομένων στον χώρο της πληροφορικής. Για παράδειγμα, κάθε φορά που κάνουμε αναζήτηση στο διαδίκτυο, συνδεόμαστε σε έναν λογαριασμό ή ολοκληρώνουμε μια συναλλαγή, το σύστημα βάσεων δεδομένων αποθηκεύει αυτές τις πληροφορίες για μελλοντική πρόσβαση.

Οι σχεσιακές βάσεις δεδομένων αποθηκεύουν τα δεδομένα σε ξεχωριστούς πίνακες, αντί να τα συγκεντρώνουν σε ένα μεγάλο αποθηκευτικό χώρο. Η δομή της βάσης δεδομένων οργανώνεται σε φυσικά αρχεία που έχουν βελτιστοποιηθεί για την ταχύτητα. Το λογικό μοντέλο των δεδομένων περιλαμβάνει διάφορες δομές, όπως πίνακες δεδομένων, προβολές, γραμμές και στήλες, και παρέχει ένα ευέλικτο περιβάλλον προγραμματισμού.

Το λογικό μοντέλο επιβάλλει κανόνες για τις σχέσεις μεταξύ διαφορετικών πεδίων δεδομένων, όπως αντιστοιχίες "ένα προς ένα", "ένα προς πολλά", μοναδικά, υποχρεωτικά ή προαιρετικά και "δείκτες" που συνδέουν διάφορους πίνακες. Αυτό διασφαλίζει ότι οι εφαρμογές που χρησιμοποιούν μια καλά σχεδιασμένη βάση δεδομένων δεν αντιμετωπίζουν ποτέ ασυνέπειες, διπλές καταχωρήσεις, απομονωμένα δεδομένα, λάθη ή ελλείποντα δεδομένα.

Το "SQL" στη "MySQL" αντιπροσωπεύει τη φράση "Structured Query Language", που είναι η κυρίαρχη γλώσσα που χρησιμοποιείται για την διαχείριση των δεδομένων στις βάσεις δεδομένων. Ανάλογα με το περιβάλλον προγραμματισμού, η SQL μπορεί να χρησιμοποιηθεί απευθείας στη διεπαφή της βάσης δεδομένων, να ενσωματωθεί σε κώδικα γραμμένο σε άλλες γλώσσες προγραμματισμού ή να χρησιμοποιηθεί με ειδικές γλωσσικές διεπαφές API που απλοποιούν τη σύνταξή της. [Ora23b][B23c].

6.2 MySQL και Workbench

Η MySQL είναι η πιο δημοφιλής βάση δεδομένων ανοιχτού κώδικα στον κόσμο. Ανοιχτός κώδικας σημαίνει ότι είναι δυνατό για οποιονδήποτε να χρησιμοποιήσει και να τροποποιήσει το λογισμικό. Οποιοσδήποτε μπορεί να κατεβάσει το λογισμικό MySQL από το Διαδίκτυο και να το χρησιμοποιήσει χωρίς να πληρώσει για αυτό. Το λογισμικό MySQL χρησιμοποιεί τη Γενική Δημόσια Άδεια GNU (GNU

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

General Public License, GPL) για να καθορίσει τι μπορεί να γίνει και τι όχι με το λογισμικό σε διαφορετικές καταστάσεις.

Σύμφωνα με την DB-Engines, η MySQL κατατάσσεται ως η δεύτερη πιο δημοφιλής βάση δεδομένων, πίσω από την Oracle Database. Η MySQL τροφοδοτεί πολλές από τις πιο προσβάσιμες εφαρμογές, συμπεριλαμβανομένων των Facebook, Twitter, Netflix, Uber, Airbnb, Shopify και Booking.com. Δεδομένου ότι η MySQL είναι ανοιχτού κώδικα, περιλαμβάνει πολλές δυνατότητες που αναπτύχθηκαν σε στενή συνεργασία με τους χρήστες για περισσότερα από 25 χρόνια. Επομένως, είναι πολύ πιθανό η εφαρμογή ή η γλώσσα προγραμματισμού να υποστηρίζεται από τη βάση δεδομένων MySQL [Ora23b].

Η MySQL είναι μία από τις πιο ευρέως χρησιμοποιούμενες βάσεις δεδομένων. Οι οργανισμοί βασίζονται στη MySQL για τη λειτουργία κρίσιμων για τις επιχειρήσεις εφαρμογών λόγω της αξιοπιστίας της. Η MySQL κλιμακώνεται ώστε να ανταποκρίνεται στις απαιτήσεις των εφαρμογών με την μεγαλύτερη πρόσβαση. Η εγγενής αρχιτεκτονική αντιγραφής της MySQL επιτρέπει σε οργανισμούς, όπως το Facebook, να κλιμακώνουν εφαρμογές για την υποστήριξη δισεκατομμυρίων χρηστών. Η MySQL παρέχει ένα πλήρες σύνολο εγγενών, πλήρως ενσωματωμένων τεχνολογιών αντιγραφής για υψηλή διαθεσιμότητα και ανάκαμψη από καταστροφές.

Για κρίσιμες εφαρμογές για τις επιχειρήσεις και για την τήρηση των δεσμεύσεων των συμφωνιών επιπέδου υπηρεσιών, οι πελάτες μπορούν να επιτύχουν. Η ασφάλεια των δεδομένων συνεπάγεται προστασία και συμμόρφωση με βιομηχανικούς και κυβερνητικούς κανονισμούς, όπως ο Γενικός Κανονισμός της Ευρωπαϊκής Ένωσης για την Προστασία Δεδομένων, 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 παρέχει προηγμένα χαρακτηριστικά ασφαλείας, όπως έλεγχο ταυτότητας/εξουσιοδότηση, διαφανή κρυπτογράφηση δεδομένων, έλεγχο, κάλυψη δεδομένων και τείχος προστασίας βάσεων δεδομένων.

Η MySQL είναι γρήγορη, αξιόπιστη, επεκτάσιμη και εύκολη στη χρήση, για τον λόγο αυτό χρησιμοποιήθηκε στην συγκεκριμένη εργασία. Αναπτύχθηκε αρχικά για να χειρίζεται γρήγορα μεγάλες βάσεις δεδομένων και έχει χρησιμοποιηθεί σε

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

εξαιρετικά απαιτητικά περιβάλλοντα παραγωγής για πολλά χρόνια. Αν και η MySQL βρίσκεται υπό συνεχή ανάπτυξη, προσφέρει ένα πλούσιο και χρήσιμο σύνολο λειτουργιών. Η συνδεσιμότητα, η ταχύτητα και η ασφάλεια της MySQL την καθιστούν εξαιρετικά κατάλληλη για πρόσβαση σε βάσεις δεδομένων στο Διαδίκτυο [Ora23b].

Πρέπει να σημειωθεί πως για την χρήση και οπτικοποίηση της MySQL απαραίτητη είναι και η χρήση του Workbench. Το MySQL Workbench είναι ένα ενοποιημένο οπτικό εργαλείο για αρχιτέκτονες βάσεων δεδομένων, προγραμματιστές και DBAs (DataBase Administrators). Το MySQL Workbench παρέχει μοντελοποίηση δεδομένων, ανάπτυξη SQL και ολοκληρωμένα εργαλεία διαχείρισης για τη διαμόρφωση του διακομιστή, τη διαχείριση χρηστών, τη δημιουργία αντιγράφων ασφαλείας και πολλά άλλα.

Το MySQL Workbench επιτρέπει σε έναν DBA, προγραμματιστή ή αρχιτέκτονα δεδομένων να σχεδιάζει, να μοντελοποιεί, να δημιουργεί και να διαχειρίζεται οπτικά βάσεις δεδομένων. Περιλαμβάνει όλα όσα χρειάζεται ένας μοντελοποιητής δεδομένων για τη δημιουργία σύνθετων μοντέλων ER, την εμπρόσθια και αντίστροφη μηχανική και παρέχει επίσης βασικά χαρακτηριστικά για την εκτέλεση δύσκολων εργασιών διαχείρισης αλλαγών και τεκμηρίωσης που συνήθως απαιτούν πολύ χρόνο και προσπάθεια. Το MySQL Workbench παρέχει οπτικά εργαλεία για τη δημιουργία, εκτέλεση και βελτιστοποίηση ερωτημάτων SQL. Ο SQL Editor παρέχει έγχρωμη επισήμανση σύνταξης, αυτόματη συμπλήρωση, επαναχρησιμοποίηση αποσπασμάτων SQL και ιστορικό εκτέλεσης της SQL.

Επιπλέον, το MySQL Workbench παρέχει μια οπτική κονσόλα για την εύκολη διαχείριση περιβαλλόντων MySQL και την καλύτερη ορατότητα των βάσεων δεδομένων. Οι προγραμματιστές και οι DBA μπορούν να χρησιμοποιούν τα οπτικά εργαλεία για τη διαμόρφωση διακομιστών, τη διαχείριση χρηστών, την εκτέλεση αντιγράφων ασφαλείας και ανάκτησης, την επιθεώρηση δεδομένων ελέγχου και την προβολή της κατάστασης της βάσης δεδομένων [Ora23c].

6.3 Διαδικασία δημιουργίας και σύνδεσης της βάσης με τον κώδικα

Για τη λήψη και αποθήκευση των ειδήσεων, απαιτείται η δημιουργία μιας βάσης δεδομένων, καθώς ο αριθμός των γεγονότων που συλλέγονται καθημερινά από την πτυχιακή εργασία είναι σημαντικός και αυξάνεται συνεχώς. Αυτό οδηγεί σε

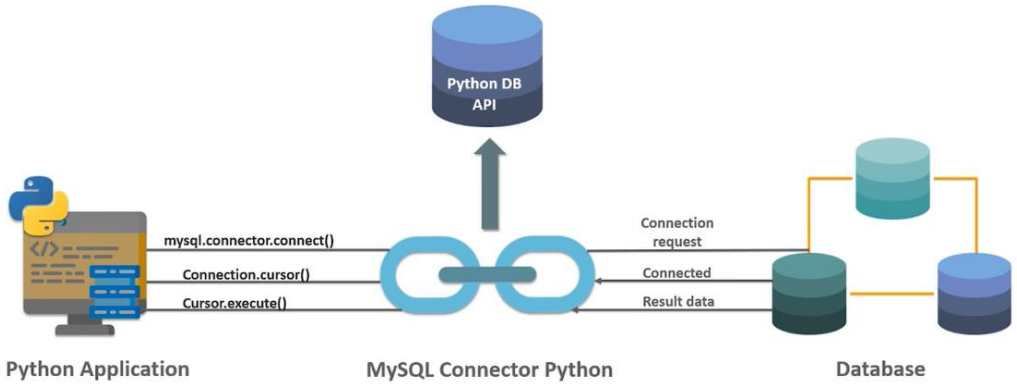
Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

καλύτερη δομή και διαχείριση των δεδομένων μέσω πινάκων. Η MySQL χρησιμοποιείται για να υλοποιήσει τη βάση δεδομένων.

Προκειμένου να κατανοηθεί η λογική πίσω από τη σύνδεση του Python κώδικα με τη βάση δεδομένων MySQL, παρουσιάζεται ένα γραφικό σχήμα (Εικόνα 38) που εξηγεί τη λειτουργία του τμήματος αυτού. Στο σχήμα, η βάση δεδομένων εμφανίζεται στα αριστερά, η εφαρμογή σε Python στα δεξιά, και στο κέντρο βρίσκεται ο συνδέτης της MySQL με την Python. Αυτός ο συνδέτης λειτουργεί ως γέφυρα μεταξύ της εφαρμογής σε Python και του διακομιστή βάσης δεδομένων, επιτρέποντας την αποτελεσματική επικοινωνία μεταξύ των δύο.

Συγκεκριμένα, η διαδικασία συνδέεται μεταξύ της εφαρμογής front-end σε Python και της βάσης δεδομένων MySQL. Αρχικά, η εφαρμογή σε Python αποστέλλει ένα αίτημα σύνδεσης (mysql.connector.connect()) στον συνδέτη. Στη συνέχεια, ο συνδέτης μεταβιβάζει το ίδιο αίτημα σύνδεσης (Connection request) στη βάση δεδομένων. Αφού η βάση δεδομένων δέχεται το αίτημα, αποστέλλει ένα μήνυμα σύνδεσης (Connected) πίσω στον συνδέτη. Στη συνέχεια, ο συνδέτης αποστέλλει ένα αίτημα δρομέα σύνδεσης (Connection.cursor()) προς τη βάση δεδομένων.

Η σύνδεση ως μέθοδος και ο δρομέας ως αντικείμενο επιτρέπουν την επικοινωνία με τη βάση δεδομένων MySQL και τη δημιουργία της δικής μας βάσης δεδομένων. Τέλος, η εφαρμογή σε Python εκτελεί μια συγκεκριμένη δήλωση συνέχειας ή ένα συγκεκριμένο ερώτημα συνέχειας (Cursor.execute()) προς τον συνδέτη, και ο διακομιστής της βάσης δεδομένων ανακτά τα δεδομένα αποτελέσματος (Result data) για χρήση στην εφαρμογή.



Εικόνα 6-1: Σύνδεση της Python με την Βάση Δεδομένων MySQL. Πηγή: https://www.youtube.com/watch?v=g60QghtJmjY

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Είναι επίσης αξιόλογο να εξετάσουμε τις βασικές λειτουργίες που μπορούν να εκτελεστούν μεταξύ της Python και της MySQL, που συνοψίζονται με το ακρωνύμιο C.R.U.D. (Create, Read, Update, Delete). Συγκεκριμένα, αυτές οι λειτουργίες επιτρέπουν τη δημιουργία, την ανάγνωση, την ενημέρωση και τη διαγραφή δεδομένων στη βάση δεδομένων MySQL μέσω της Python.

Για να διευκολυνθεί τη διαχείριση των δεδομένων, η δημιουργία μιας βάσης δεδομένων αποτελεί κρίσιμο στάδιο. Η Python, ως ευρέως χρησιμοποιούμενη γλώσσα προγραμματισμού στην επιστημονική κοινότητα, αποτελεί ιδανικό εργαλείο για τη διαχείριση των δεδομένων. Όμως, η ανάπτυξη μιας εφαρμογής θα ήταν αδύνατη χωρίς τη γνώση του τόπου αποθήκευσης των δεδομένων. Εδώ είναι που ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS), όπως ο MySQL DB Server, καθίσταται κρίσιμο για την αποτελεσματική ανάπτυξη εφαρμογών.

6.3.1 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.

```
# Εισαγωγή των στοιχείων του total_crime_accident_df στη βάση δεδομένων greekeventdb
    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 6-2: 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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

- ¨**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.

6.3.2 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 6-3: Δημιουργία μιας μηχανής (engine) βάσης δεδομένων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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/greek eventdb': 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 συνθεθεί 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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

6.4 Επίλογος

Η MySQL αποτελεί θεμέλιο στο συνεχώς εξελισσόμενο τοπίο της διαχείρισης δεδομένων. Η ευρεία υιοθέτησή της σε ποικίλες εφαρμογές και η χρήση της από τεχνολογικούς κολοσσούς αποδεικνύουν την ευελιξία και την αξιοπιστία της. Η ευελιξία και η φιλικότητα προς τον χρήστη της MySQL είναι μερικοί από τους λόγους της δημοτικότητάς της.

Σύμφωνα, λοιπόν. με τον κώδικα που αναλύθηκε παραπάνω για την συγκεκριμένη εργασία, αναφέρθηκε η σύνδεση σε στη βάση δεδομένων MySQL χρησιμοποιώντας τη βιβλιοθήκη **mysql.connector** και τη συνάρτηση **create_engine** από τη βιβλιοθήκη SQLAlchemy. Η δημιουργία των πινάκων στη βάση δεδομένων, όπου αποθηκεύονται τα ειδησεογραφικά γεγονότα και η διαδικασία προετοιμασίας και αποθήκευσής τους πραγματεύονται στις Ενότητες (6.3.3.3) έως (6.3.3.5) της πτυχιακής εργασίας [Σ23].

Επιπλέον, είναι σημαντικό να αναφερθεί πως η ιδιότητα της MySQL ως βιομηχανικό πρότυπο έχει προωθήσει μια πλούσια δεξαμενή πόρων και εξειδικευμένων προγραμματιστών, εξασφαλίζοντας συνεχή ανάπτυξη και άφθονη υποστήριξη για τους χρήστες. Παραμένει ένα ανεκτίμητο εργαλείο για την αξιοποίηση του δυναμικού των δεδομένων με αποτελεσματικότητα, απόδοση και ασφάλεια. Η εκμάθηση της MySQL ανοίγει τις πόρτες σε έναν κόσμο δυνατοτήτων βασισμένων στα δεδομένα, δίνοντάς τη δυνατότητα της διαρκής εξέλιξης του τοπίου της τεχνολογίας και των επιχειρήσεων [B23c].

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

7 Εντοπισμός πανομοιότυπων ή σχεδόν όμοιων ειδησεογραφικών άρθρων

7.1 Φιλτράρισμα πανομοιότυπων ή σχεδόν όμοιων ειδήσεων στο DataFrame βάσει τίτλου και περιγραφής

7.1.1 Εισαγωγή

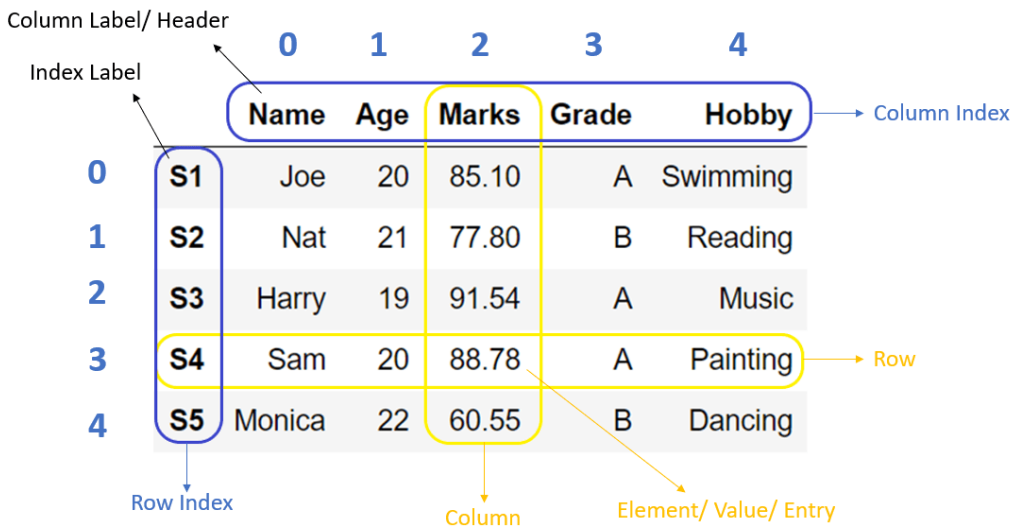
Η διαγραφή σχεδόν διπλότυπων ειδήσεων είναι ζωτικής σημασίας για τη διατήρηση της ποιότητας, της αποτελεσματικότητας και της χρηστικότητας μιας βάσης δεδομένων που περιέχει γεγονότα άρθρων ειδήσεων. Τα σχεδόν διπλότυπα άρθρα περιέχουν σχεδόν πανομοιότυπο περιεχόμενο, που διαφέρουν μόνο σε μικρές παραλλαγές. Η αποθήκευση πολλών αντιγράφων τέτοιων άρθρων διογκώνει άσκοπα το μέγεθος της βάσης δεδομένων και οδηγεί σε πλεονασμό δεδομένων. Καταργώντας σχεδόν διπλότυπα, εξοικονομείται χώρος αποθήκευσης και να βελτιστοποιείται η απόδοση της βάσης δεδομένων.

Η κατάργηση σχεδόν διπλότυπων διασφαλίζει ότι οι χρήστες παρουσιάζονται με ένα πιο συνοπτικό και συνεκτικό σύνολο αποτελεσμάτων αναζήτησης, βελτιώνοντας την εμπειρία τους και δίνοντάς τους τη δυνατότητα να έχουν πρόσβαση στο πιο σχετικό περιεχόμενο. Επιπλέον, η διαχείριση μιας βάσης δεδομένων με μεγάλο αριθμό σχεδόν διπλότυπων μπορεί να καταστήσει πιο περίπλοκες και χρονοβόρες τις εργασίες συντήρησης της.

Πριν αναλύσουμε τις μεθόδους που χρησιμοποιήθηκαν, θα αναφερθούμε συντόμως στην έννοια των DataFrame, διότι είναι η δομή στην οποία εξετάζεται ο έλεγχος διπλοτύπων αυτού του υποκεφαλαίου. 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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Ένα DataFrame αντιπροσωπεύει μια δομή δεδομένων που συγκρίνεται ευλόγως με έναν πίνακα, παρόμοιο με τον τρόπο που θα αναφερόταν σε έναν πίνακα SQL ή ένα υπολογιστικό φύλλο. Στην ουσία του, αυτός ο πίνακας αποτελείται από σειρές και στήλες, όπου κάθε στήλη μπορεί να περιλαμβάνει δεδομένα διάφορων τύπων, όπως αριθμητικά στοιχεία, συμβολοσειρές και άλλα. Τα DataFrames επιτρέπουν την αποθήκευση και τη διαχείριση δομημένων δεδομένων με αποτελεσματικό τρόπο.



Εικόνα 7-1: Παράδειγμα μορφής DataFrame που περιέχει ανθρώπους και ορισμένα χαρακτηριστικά τους. Πήγη: pynative.com

Για τη δημιουργία ενός Pandas DataFrame, μπορούμε να χρησιμοποιήσουμε λίστες, λεξικά ή ακόμη και λίστες από λεξικά. Τα DataFrames παρέχουν πληθώρα λειτουργιών, όπως φιλτράρισμα, συγχώνευση, ομαδοποίηση και άλλες, καθιστώντας εφικτή την αποτελεσματική εκτέλεση διάφορων λειτουργιών ανάλυσης δεδομένων. Επιπλέον, διαχειρίζονται εύκολα τα απουσιάζοντα δεδομένα και αντιμετωπίζουν διάφορους τύπους δεδομένων εντός της ίδιας δομής, γεγονός που τα καθιστά εξαιρετικά ευέλικτα για ποικίλες εργασίες χειρισμού δεδομένων.

Χρησιμοποιήθηκαν τρεις βασικές μέθοδοι για την ανίχνευση και τη διαγραφή των σχεδόν διπλοτύπων άρθρων ειδήσεων για να εξασφαλιστεί όσο είναι δυνατόν η αποτελεσματικότητα της μη ύπαρξης διπλότυπων στη βάση δεδομένων.

7.1.2 Πρώτη μέθοδος

Ακολουθεί η συνάρτηση **similarity_filter**, η οποία πραγματοποιεί την εύρεση και τη αφαίρεση σχεδόν διπλότυπων άρθρων ειδήσεων με έλεγχο του τίτλου τους,

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

χρησιμοποιώντας ένα μέτρο ομοιότητας που παρέχεται από τη μέθοδο .similarity() της βιβλιοθήκης NLP, spaCy. Υπολογίζει τυπικά την ομοιότητα συνημιτόνου μεταξύ των διανυσματικών αναπαραστάσεων των δύο εγγράφων.

```
# Αναδρομικός έλεγχος αν οι τίτλοι περνούν από το similarity filter
def similarity_filter(titles, limit):

    # Προ-επεξεργασία τίτλων
    preprocessed_title_docs = pre_process(titles)

    # Αφαίρεση παρόμοιων τίτλων
    all_summary_pairs = list(combinations(preprocessed_title_docs, 2))
    similar_titles = []
    for pair in all_summary_pairs:
        title1 = nlp(pair[0])
        title2 = nlp(pair[1])
        similarity = title1.similarity(title2)
        if similarity > limit:
            similar_titles.append(pair)

    titles_to_remove = []
    for a_title in similar_titles:
        # Λήψη του index του πρώτου τίτλου στο ζευγάρι
        index_for_removal = preprocessed_title_docs.index(a_title[0])
        titles_to_remove.append(index_for_removal)

    # Λήψη των index των παρόμοιων ειδήσεων και αφαίρεση τους
    similar_title_counts = set(titles_to_remove)

    similar_titles = [
        x[1] for x in enumerate(titles) if x[0] in similar_title_counts
    ]

    # Έξοδος από την αναδρομή αν δεν υπάρχουν άλλοι παρόμοιοι τίτλοι
    if len(similar_title_counts) == 0:
        return titles

    # Συνέχεια της αναδρομής αν υπάρχουν ακόμα παρόμοιοι τίτλοι
    else:
        # Αφαίρεση παρόμοιων τίτλων από την επόμενη είσοδο
        for title in similar_titles:
            idx = titles.index(title)
            duplicate_list.append(title)
            titles.pop(idx)

        return similarity_filter(titles, limit)
```

Εικόνα 7-2: Συνάρτηση similarity_filter.

Αυτό το απόσπασμα κώδικα εφαρμόζει μια αναδρομική συνάρτηση φίλτρου ομοιότητας σε μια λίστα τίτλων. Ο στόχος της συνάρτησης είναι να αφαιρέσει τίτλους

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

που θεωρούνται πολύ όμοιοι μεταξύ τους, όπως καθορίζεται από ένα όριο ομοιότητας (limit). Ακολουθεί μια ανάλυση του κώδικα:

1. Η συνάρτηση ξεκινά με προεπεξεργασία των τίτλων εισόδου χρησιμοποιώντας μια συνάρτηση **pre_process** που αναλύεται στην Ενότητα (6.2.2) της εργασίας [Σ23]. Αυτό το βήμα περιλαμβάνει καθαρισμό και τυποποίηση των τίτλων για σύγκριση.
2. Στη συνέχεια, η συνάρτηση δημιουργεί ζεύγη τίτλων χρησιμοποιώντας το «combinations» από τη μονάδα itertools και υπολογίζει την ομοιότητα μεταξύ κάθε ζεύγους χρησιμοποιώντας ένα μοντέλο spaCy NLP (nlp) με την γραμμή κώδικα: *similarity = title1.similarity(title2).*
    - title1 και title2: Πρόκειται για αντικείμενα που αντιπροσωπεύουν τίτλους άρθρων ειδήσεων.
    - .similarity(title2): Αυτή είναι μια κλήση μεθόδου που υπολογίζει την ομοιότητα μεταξύ των δύο εγγράφων κειμένου με βάση το περιεχόμενό τους. Ο συγκεκριμένος υπολογισμός της ομοιότητας μπορεί να ποικίλλει ανάλογα με τη βιβλιοθήκη NLP που χρησιμοποιείται. Στο spaCy, η μέθοδος .similarity() υπολογίζει τυπικά την ομοιότητα συνημιτόνου μεταξύ των διανυσματικών αναπαραστάσεων των δύο εγγράφων.
    - similarity: Αυτή η μεταβλητή διατηρεί την υπολογιζόμενη βαθμολογία ομοιότητας, η οποία είναι μια τιμή κινητής υποδιαστολής μεταξύ 0 και 1. Μια υψηλότερη βαθμολογία ομοιότητας υποδηλώνει ότι τα δύο έγγραφα μοιάζουν περισσότερο ως προς το περιεχόμενο και τη δομή τους.
3. Εάν η ομοιότητα μεταξύ δύο τίτλων υπερβαίνει το δεδομένο όριο, το ζεύγος τίτλων θεωρείται παρόμοιο και προστίθεται στη λίστα **similar_titles**.
4. Για κάθε παρόμοιο ζεύγος τίτλων, το ευρετήριο του πρώτου τίτλου στο ζεύγος αποθηκεύεται στη λίστα **titles_to_remove**.
5. Στη συνέχεια, ο κώδικας δημιουργεί ένα σύνολο δεικτών που πρέπει να αφαιρεθούν **similar_title_counts**, που αντιστοιχούν στους τίτλους που έχουν επισημανθεί για διαγραφή.
6. Εάν δεν βρέθηκαν παρόμοιοι τίτλοι δηλαδή, το **similar_title_counts** είναι κενό, η συνάρτηση επιστρέφει τους αρχικούς τίτλους.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

7. Εάν βρέθηκαν παρόμοιοι τίτλοι, γίνεται μια αναδρομική κλήση στη συνάρτηση φίλτρου-ομοιότητας μετά την αφαίρεση των παρόμοιων τίτλων. Αυτό συνεχίζεται μέχρι να μην απομείνουν άλλοι παρόμοιοι τίτλοι.

Ο γενικός σκοπός αυτής της αναδρομικής συνάρτησης είναι η επαναληπτική κατάργηση τίτλων που είναι πολύ παρόμοιοι μεταξύ τους, με τη διαδικασία να συνεχίζεται μέχρι να μην βρεθούν άλλοι παρόμοιοι τίτλοι εντός του καθορισμένου ορίου ομοιότητας. Είναι σημαντικό να διασφαλιστεί ότι το όριο ομοιότητας έχει επιλεγεί κατάλληλα για να επιτευχθεί το επιθυμητό επίπεδο διαφοροποίησης μεταξύ των τίτλων.

7.1.3 Δεύτερη μέθοδος

Ακολουθεί η συνάρτηση **remove_duplicates**, η οποία πραγματοποιεί την εύρεση και τη αφαίρεση σχεδόν διπλότυπων άρθρων ειδήσεων με έλεγχο του τίτλου και της περιγραφής τους, χρησιμοποιώντας sentence embeddings.

```
# Αφαίρεση παρόμοιων ειδήσεων (άλλη μέθοδος)
def remove_duplicates(text_list, similarity_threshold):
    similar_pairs = []
    for pair in itertools.combinations(text_list, 2):
        text1_embedding = model.encode(pair[0])
        text2_embedding = model.encode(pair[1])
        similarity = 1 - scipy.spatial.distance.cosine(text1_embedding, text2_embedding)
        if similarity >= similarity_threshold:
            similar_pairs.append((pair[0], pair[1], similarity))
    for i, text in enumerate(text_list):
        is_duplicate = False
        for pair in similar_pairs:
            if text in pair:
                is_duplicate = True
                break
        if is_duplicate:
            duplicate_list.append(text_list[i])
            del text_list[i]
    return text_list
```

Εικόνα 7-3: Συνάρτηση remove_duplicates.

Αυτό το απόσπασμα κώδικα ορίζει μια συνάρτηση **remove_duplicates** που αφαιρεί παρόμοια στοιχεία από μια λίστα με βάση ένα όριο ομοιότητας. Επικεντρώνεται στη σύγκριση του σημασιολογικού περιεχομένου των στοιχείων

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

κειμένου χρησιμοποιώντας embeddings από ένα μοντέλο embedding «sentence embeddings» και στη συνέχεια στην διαγραφή στοιχείων που θεωρούνται πολύ παρόμοια. Τα sentence embeddings είναι αριθμητικές αναπαραστάσεις προτάσεων που αποτυπώνουν τη σημασιολογική τους σημασία, επιτρέποντας ουσιαστικές συγκρίσεις μεταξύ των προτάσεων. Ακολουθεί μια ανάλυση του κώδικα:

Ο κώδικας προϋποθέτει την ύπαρξη ενός προ-εκπαιδευμένου μοντέλου ενσωμάτωσης προτάσεων, που αναγράφεται ως model. Αυτό το μοντέλο είναι ικανό να κωδικοποιεί προτάσεις σε πυκνά αριθμητικά διανύσματα που συλλαμβάνουν τις σημασιολογικές τους πληροφορίες. Επιπλέον:

- **text_list**: Μια λίστα στοιχείων κειμένου (π.χ. τίτλοι άρθρων ειδήσεων ή περιγραφές) που πρόκειται να υποβληθούν σε επεξεργασία για διπλότυπη διαγραφή.
- **similarity_threshold**: Μια τιμή κατωφλίου που καθορίζει το επίπεδο ομοιότητας που απαιτείται για να θεωρηθούν διπλότυπα δύο στοιχεία κειμένου.

1. Ο κώδικας επαναλαμβάνεται σε όλα τα ζεύγη στοιχείων κειμένου χρησιμοποιώντας το itertools.combinations. Για κάθε ζεύγος στοιχείων κειμένου στη λίστα κειμένου, ο κώδικας υπολογίζει τη σημασιολογική τους ομοιότητα χρησιμοποιώντας sentence embeddings. Αυτό γίνεται με την κωδικοποίηση κάθε στοιχείου κειμένου χρησιμοποιώντας τη μέθοδο model.encode(). Οι προκύπτουσες ενσωματώσεις χρησιμοποιούνται στη συνέχεια για τον υπολογισμό της ομοιότητας συνημιτόνου μεταξύ των δύο sentence embeddings.
2. Η απόσταση συνημιτόνου μεταξύ των ενσωματώσεων υπολογίζεται χρησιμοποιώντας το **scipy.spatial.distance.cosine** και, στη συνέχεια, η ομοιότητα προκύπτει ως 1 - cosine_distance.
3. Εάν η υπολογισμένη ομοιότητα είναι μεγαλύτερη ή ίση με το καθορισμένο όριο, το ζεύγος παρόμοιων στοιχείων κειμένου και η βαθμολογία ομοιότητάς τους προστίθενται στη λίστα **similar_pairs**.
4. Στη συνέχεια, ο κώδικας περνά μέσα από την αρχική λίστα κειμένου και ελέγχει εάν κάθε στοιχείο κειμένου υπάρχει σε κάποιο από τα παρόμοια ζεύγη. Εάν ναι, επισημαίνει το στοιχείο ως διπλότυπο προσαρτώντας το στη λίστα **duplicate_list** και το αφαιρεί από τη λίστα κειμένου.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

5. Μετά την επεξεργασία όλων των στοιχείων, η συνάρτηση επιστρέφει την τροποποιημένη λίστα κειμένου, τώρα χωρίς τα διπλότυπα που έχουν αφαιρεθεί.

Αυτή η συνάρτηση αξιοποιεί αποτελεσματικά τη σημασιολογική ομοιότητα για τον εντοπισμό και την αφαίρεση σχεδόν διπλότυπων στοιχείων κειμένου από μια λίστα, παρέχοντας μια εναλλακτική προσέγγιση για την ανίχνευση διπλότυπων. Η επιλογή του ορίου ομοιότητας είναι σημαντική, καθώς καθορίζει το επίπεδο ευαισθησίας στην ανίχνευση σχεδόν διπλότυπων. Αξίζει να σημειωθεί ότι το απόσπασμα κώδικα προϋποθέτει τη σωστή διαμόρφωση τόσο του μοντέλου sentence embeddings όσο και του υπολογισμού απόστασης scipy.spatial.distance.cosine.

7.1.4 Τρίτη μέθοδος

Ακολουθεί η συνάρτηση **near_duplicates**, η οποία πραγματοποιεί την εύρεση σχεδόν διπλότυπων άρθρων ειδήσεων με έλεγχο του τίτλου τους, χρησιμοποιώντας ένα μέτρο ομοιότητας που παρέχεται από τη συνάρτηση similarity(a,b). Επιπλέον, δημιουργεί και αποθηκεύει τα διπλότυπα που ανιχνεύει σε ένα Dataframe **df_near_duplicates** για περαιτέρω ανάλυση και έλεγχο σωστής λειτουργίας της συνάρτησης.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

```python
def near_duplicates(df, duplicate_df):
    def similarity(a, b):
        return SequenceMatcher(None, a, b).ratio()

    # Υπολογισμός της ομοιότητας ανά ζεύγη μεταξύ των τίτλων
    similarity_matrix = [[similarity(a, b) for b in df['title']] for a in df['title']]

    # Δημιουργία νέου DataFrame για την αποθήκευση των ομοιοτήτων κατά ζεύγη
    similarity_df = pd.DataFrame(similarity_matrix, columns=df.index, index=df.index)

    # Φιλτράρισμα του DataFrame για διατήρηση μόνο των σχεδόν διπλοτύπων με similarity score πάνω από ένα threshold
    threshold = 0.8
    near_duplicates = (similarity_df > threshold).sum() - 1
    near_duplicates = near_duplicates[near_duplicates > 0].index.tolist()
    df_near_duplicates = df.loc[near_duplicates]

    # Διατήρηση της πρώτης σειράς για κάθε σύνολο σχεδόν διπλότυπων
    df_near_duplicates2 = df_near_duplicates.drop_duplicates(subset='title', keep='first')
    left_join = pd.merge(df_near_duplicates2, df_near_duplicates, on='title', how='left')
    duplicate_df = pd.concat([duplicate_df, df_near_duplicates])
    # print("left join: ", left_join)

    # Λήψη της λίστας με τις τιμές index των σχεδόν διπλότυπων
    index_near_duplicates = df_near_duplicates2.index.tolist()

    # Λήψη της λίστας με τις τιμές index για τους τίτλους που δεν έχουν σχεδόν διπλότυπα
    index_no_near_duplicates = [idx for idx in df.index.tolist() if idx not in index_near_duplicates]

    # Λήψη του DataFrame των τίτλων που δεν έχουν σχεδόν διπλότυπα
    df_no_near_duplicates = df.loc[index_no_near_duplicates]

    # Συνένωση των DataFrames
    df = pd.concat([df_near_duplicates2, df_no_near_duplicates])
    #print(df)
```

Εικόνα 7-4: Η συνάρτηση near_duplicates.

Ακολουθεί η ανάλυση του τρόπου λειτουργίας της συνάρτησης.

1. Η συνάρτηση ορίζει μια συνάρτηση υπολογισμού ομοιότητας **similarity(a,b)** που υπολογίζει την αναλογία ομοιότητας μεταξύ δύο συμβολοσειρών κειμένου a και b, οι οποίες αντιπροσωπεύουν τους τίτλους δύο άρθρων ειδήσεων. Χρησιμοποιεί την κλάση SequenceMatcher από τη μονάδα difflib για να υπολογίσει την αναλογία ομοιότητας μεταξύ αυτών των δύο συμβολοσειρών. Η μέθοδος ratio() της κλάσης SequenceMatcher υπολογίζει ένα σκορ ομοιότητας που κυμαίνεται από 0 έως 1, όπου το 1 υποδηλώνει ότι οι συμβολοσειρές είναι ίδιες και το 0 δεν υποδηλώνει καμία ομοιότητα.

   Για παράδειγμα, εάν το a είναι "Ειδήσεις: Ο σεισμός χτυπά την πόλη" και το b είναι "Πόλη που χτυπήθηκε από καταστροφικό σεισμό", η συνάρτηση ομοιότητας θα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

επιστρέψει ένα σκορ ομοιότητας με βάση το πόσο όμοιες είναι οι δύο συμβολοσειρές από την άποψη της ακολουθίας χαρακτήρων.

2. Η συνάρτηση δημιουργεί έναν πίνακα ομοιότητας υπολογίζοντας την ομοιότητα μεταξύ ζευγαριών τίτλων άρθρων ειδήσεων. Κάθε κελί στη μήτρα αντιπροσωπεύει τη βαθμολογία ομοιότητας μεταξύ δύο τίτλων. Ο εξωτερικός βρόχος επαναλαμβάνεται πάνω από κάθε τίτλο a στο Dataframe **df** και ο εσωτερικός βρόχος επαναλαμβάνεται πάνω από κάθε τίτλο b στο Dataframe **df.** Για κάθε ζεύγος τίτλων (a, b), η συνάρτηση ομοιότητας καλείται να υπολογίσει τη βαθμολογία ομοιότητας και αυτή η βαθμολογία αποθηκεύεται στο αντίστοιχο κελί του πίνακα ομοιότητας.

   Το αποτέλεσμα είναι ένας δισδιάστατος πίνακας όπου κάθε γραμμή και στήλη αντιστοιχεί σε έναν τίτλο ειδησεογραφικού άρθρου και κάθε κελί περιέχει τη βαθμολογία ομοιότητας μεταξύ των τίτλων των αντίστοιχων άρθρων. Ο πίνακας καταγράφει τον βαθμό ομοιότητας μεταξύ όλων των ζευγών τίτλων ειδήσεων. Αυτός ο πίνακας ομοιότητας χρησιμοποιείται αργότερα για τον εντοπισμό σχεδόν διπλότυπων ειδήσεων συγκρίνοντας τις βαθμολογίες ομοιότητας έναντι ενός προκαθορισμένου ορίου (threshold).

3. Ο πίνακας ομοιότητας χρησιμοποιείται για τη δημιουργία ενός Dataframe **similarity_df** όπου οι γραμμές και οι στήλες αντιστοιχούν στους δείκτες των άρθρων ειδήσεων. Κάθε κελί σε αυτό το Dataframe έχει τη βαθμολογία ομοιότητας μεταξύ των τίτλων του αντίστοιχου ζεύγους ειδησεογραφικών άρθρων.

4. Η συνάρτηση εφαρμόζει ένα όριο ομοιότητας για να μετρήσει τον αριθμό των άρθρων που έχουν βαθμολογία ομοιότητας μεγαλύτερη από το όριο για κάθε άρθρο. Τα άρθρα με τουλάχιστον ένα σχεδόν διπλότυπο (δηλαδή βαθμολογία ομοιότητας πάνω από το όριο) διατηρούνται και οι δείκτες τους αποθηκεύονται.

   - Το όριο του 0,8 επιλέχθηκε με βάση μια αντιστάθμιση μεταξύ ακρίβειας και ανάκλησης για τον εντοπισμό σχεδόν διπλότυπων. Στο πλαίσιο της ανίχνευσης σχεδόν διπλότυπων, η ακρίβεια αναφέρεται στην ακρίβεια αναγνώρισης αληθινών σχεδόν διπλότυπων και η ανάκληση αναφέρεται στην ικανότητα καταγραφής όλων των πραγματικών σχεδόν διπλότυπων που υπάρχουν στα δεδομένα.

   Η επιλογή ενός ορίου περιλαμβάνει την εξέταση των ακόλουθων παραγόντων:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

- Ακρίβεια έναντι ανάκλησης: Ένα υψηλότερο όριο, όπως το 0,8, αυξάνει την πιθανότητα τα σχεδόν διπλότυπα που ανιχνεύονται να είναι πράγματι πολύ παρόμοια. Αυτό μπορεί να οδηγήσει σε μεγαλύτερη ακρίβεια, που σημαίνει ότι τα σχεδόν διπλότυπα που προσδιορίζονται είναι πιο πιθανό να είναι αληθινά σχεδόν διπλότυπα. Ωστόσο, μπορεί να οδηγήσει σε απώλεια ορισμένων σχεδόν διπλότυπων και χαμηλότερη ανάκληση.
- Εσφαλμένα θετικά και ψευδώς αρνητικά: Ένα χαμηλότερο όριο θα ανιχνεύει περισσότερα σχεδόν διπλότυπα, αλλά μπορεί επίσης να περιλαμβάνει περισσότερα ψευδώς θετικά δηλαδή, περιπτώσεις που στην πραγματικότητα δεν είναι σχεδόν διπλότυπες. Ένα υψηλότερο όριο μειώνει τα ψευδώς θετικά, αλλά μπορεί να οδηγήσει σε ψευδώς αρνητικά, δηλαδή πραγματικά σχεδόν διπλότυπα που δεν ανιχνεύονται.

Τελικά, η τιμή κατωφλίου επιλέγεται με βάση μια προσεκτική εξέταση της αντιστάθμισης ακριβείας-ανάκλησης, των ειδικών στόχων της ανίχνευσης σχεδόν διπλοτύπων και των χαρακτηριστικών του συνόλου δεδομένων.

5. Το Dataframe **df_near_duplicates** φιλτράρεται για να περιλαμβάνει μόνο τα σχεδόν διπλότυπα άρθρα ειδήσεων με βάση τους δείκτες τους. Τα αντίγραφα διατηρούνται μόνο μία φορά, διατηρώντας την πρώτη εμφάνιση.
6. Εκτελείται LEFT JOIN στο Dataframe που περιέχει σχεδόν διπλότυπα άρθρα για τη δημιουργία ενός συγχωνευμένου Dataframe που περιλαμβάνει πληροφορίες σχετικά με τα πρωτότυπα και σχεδόν διπλότυπα άρθρα.
7. Μια λίστα που περιέχει διπλότυπα άρθρα ειδήσεων ενημερώνεται συνενώνοντάς την με το Dataframe των σχεδόν διπλότυπων άρθρων **duplicate_df**.
    - Το Dataframe των σχεδόν διπλότυπων άρθρων **duplicate_df** δημιουργείται με τον παρακάτω κώδικα:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

```
duplicate_df = pd.DataFrame(columns=['title', 'category', 'description', 'link', 'content', 'pubDate', 'sublevel'])

    for title in duplicate_list:
        matching_rows = df[df['title'].str.contains(title, regex=False)]

        if len(matching_rows) > 0:
            duplicate_df = pd.concat([duplicate_df, matching_rows])

    duplicate_df = duplicate_df.drop_duplicates(subset="title")  # αφαίρεση διπλότυπων ειδήσεων σύμφωνα με τον τίτλο
    # print('\nDuplicates Dataframe: ', duplicate_df)
```

Εικόνα 7-5: Δημιουργία Dataframe duplicate_df.

Αυτό το απόσπασμα κώδικα αποτελεί μέρος μιας διαδικασίας αναγνώρισης και συλλογής διπλότυπων ειδήσεων από ένα Dataframe **df** με βάση μια λίστα τίτλων **duplicate_list**. Στη συνέχεια, δημιουργεί ένα νέο Dataframe **duplicate_df** για την αποθήκευση αυτών των διπλότυπων άρθρων, διασφαλίζοντας ταυτόχρονα ότι τα διπλότυπα άρθρα με τον ίδιο τίτλο καταργούνται. Ακολουθεί μια ανάλυση του κώδικα:

1. Δημιουργείται ένα νέο Dataframe duplicate_df με συγκεκριμένες στήλες: 'title', 'category', 'description', 'link', 'content', 'pubDate', 'sublevel'.
2. Για κάθε τίτλο στη λίστα **duplicate_list**, ελέγχει εάν τυχόν σειρές στο αρχικό Dataframe **df** έχουν τίτλους που περιέχουν τον τρέχοντα τίτλο.
3. Εάν υπάρχουν διπλότυπα άρθρα, αυτές οι σειρές αποθηκεύονται στο **matching_rows** Dataframe.
4. Εάν υπάρχουν διπλότυπες σειρές, προσαρτά αυτές τις σειρές στο **duplicate_df** Dataframe χρησιμοποιώντας το pd.concat().
5. Αφού συλλεχθούν όλα τα διπλότυπα, ο κώδικας αφαιρεί τα διπλότυπα από το **duplicate_df** Dataframe με βάση τη στήλη 'title'. Αυτό διασφαλίζει ότι μόνο μοναδικά άρθρα με διακριτούς τίτλους διατηρούνται στο **duplicate_df.**

8. Δημιουργούνται λίστες index για σχεδόν διπλότυπα άρθρα **index_near_duplicates** και μη σχεδόν διπλότυπα άρθρα **index_no_near_duplicates**.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

9. Το Dataframe **df_no_near_duplicates** φιλτράρεται ξανά για να περιλαμβάνει μόνο τα άρθρα ειδήσεων που δεν είναι σχεδόν διπλότυπα, χρησιμοποιώντας τη λίστα ευρετηρίων για μη σχεδόν διπλότυπα.
10. Η συνάρτηση συνενώνει τα Dataframes για σχεδόν διπλότυπα άρθρα και τα μη σχεδόν διπλότυπα άρθρα για να δημιουργήσει ένα καθαρισμένο Dataframe **df** που αφαιρεί σχεδόν διπλότυπους τίτλους από τη συλλογή άρθρων ειδήσεων.

Συνοπτικά, αυτή η συνάρτηση επεξεργάζεται μια συλλογή άρθρων ειδήσεων για να εντοπίσει σχεδόν διπλότυπους τίτλους με βάση τις βαθμολογίες ομοιότητάς τους. Διατηρεί την πρώτη εμφάνιση κάθε συνόλου σχεδόν διπλότυπων και δημιουργεί ένα καθαρισμένο Dataframe που διαχωρίζει σχεδόν διπλότυπα άρθρα από μη σχεδόν διπλότυπα άρθρα. Αυτό βοηθά στη διατήρηση της ποιότητας των δεδομένων και στη διασφάλιση ότι παρόμοια άρθρα ειδήσεων αντιμετωπίζονται κατάλληλα κατά την ανάλυση.

Η παρακάτω εικόνα δείχνει τον τρόπο κλήσης της συνάρτησης near_duplicates για όλες τις κατηγορίες.

```
rows = len(df.axes[0])
  print(rows)

  if rows > 200:
      # Επιπλέον έλεγχος των dataframe για εύρεση και διαγραφή παρόμοιων ειδήσεων
      near_duplicates(total_crime_accident_df, duplicate_df)
      near_duplicates(total_environment_df, duplicate_df)
      near_duplicates(total_health_df, duplicate_df)
      near_duplicates(total_social_unrest_df, duplicate_df)
      near_duplicates(total_science_df, duplicate_df)
      near_duplicates(total_arts_df, duplicate_df)
      near_duplicates(total_olympic_games_df, duplicate_df)
      near_duplicates(total_education_df, duplicate_df)
      near_duplicates(total_religion_df, duplicate_df)
      near_duplicates(total_economy_df, duplicate_df)
      near_duplicates(total_politics_df, duplicate_df)
```

Εικόνα 7-6: Κλήση συνάρτησης near_duplicates.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

7.1.5 Επίλογος

Συνοπτικά, η κατάργηση σχεδόν διπλότυπων άρθρων από μια βάση δεδομένων ειδήσεων βελτιώνει την ποιότητα των δεδομένων, την εμπειρία χρήστη, τη συνάφεια αναζήτησης και τη χρήση των πόρων. Εξασφαλίζει ότι η βάση δεδομένων παραμένει αποτελεσματική, ακριβής και φιλική προς το χρήστη ενώ υποστηρίζει διάφορες διαδικασίες ανάλυσης και λήψης αποφάσεων.

Με τη χρήση μέτρων ομοιότητας για την εύρεση των σχεδόν διπλοτύπων, είναι σημαντική η δοκιμή και να αξιολόγηση του επιλεγμένου ορίου σε αντιπροσωπευτικά δεδομένα για να διασφαλιστεί ότι πληροί την επιθυμητή ισορροπία μεταξύ ακρίβειας και πληρότητας.

7.2 Φιλτράρισμα πανομοιότυπων ή σχεδόν όμοιων ειδήσεων στη βάση δεδομένων

7.2.1 Εισαγωγή

Αναφερόμενοι σε όλες τις παραπάνω μεθόδους αφαίρεσης διπλότυπων ή σχεδόν διπλότυπων ειδήσεων θα μπορούσε κανείς να πει πως το «πρόβλημα» αυτό έχει λυθεί. Όμως, όλοι οι παραπάνω τρόποι αναφέρονται στις ειδήσεις που βρίσκονται στο dataframe και δεν έχουν εισαχθεί ακόμα στην βάση δεδομένων. Επομένως, μέχρι τώρα έχει γίνει γνωστό πως ανά μία ώρα που τρέχει ο κώδικας, το dataframe κρατά τις ειδήσεις που είναι διαφορετικές αφαιρώντας τις διπλότυπες. Παρόλα αυτά, όταν γίνεται εισαγωγή των ειδήσεων στη βάση ανά μία ώρα, έχουν εισαχθεί και κάποιες ίδιες ή σχεδόν ίδιες ειδήσεις που επιβαρύνουν την βάση με άσκοπο χώρο αποθήκευσης.

Έτσι κρίνεται μείζον ζήτημα η χρήση μεθόδων και συναρτήσεων προκειμένου να υπάρχουν στην βάση ουσιώδης και χρήσιμες ειδήσεις μόνο μια φορά. Η κατάργηση διπλότυπων ή σχεδόν διπλότυπων από μια βάση δεδομένων ειδήσεων είναι πολύ χρήσιμη για διάφορους λόγους:

1. Τα διπλότυπα ή σχεδόν διπλότυπα άρθρα μπορούν να παραμορφώσουν τη στατιστική ανάλυση και τις πληροφορίες που προέρχονται από τη βάση δεδομένων. Καταργώντας αυτά τα διπλότυπα, διασφαλίζεται ότι τα δεδομένα αντιπροσωπεύουν με ακρίβεια ξεχωριστά κομμάτια πληροφοριών.
2. Κατά τη διεξαγωγή αναλύσεων, όπως ανάλυση συναισθήματος, μοντελοποίηση θεμάτων ή ανίχνευση τάσεων, τα διπλά άρθρα μπορούν να διογκώσουν τη σημασία ορισμένων θεμάτων ή συναισθημάτων, οδηγώντας σε παραπλανητικά

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

συμπεράσματα. Η κατάργηση των διπλότυπων διασφαλίζει ότι η ανάλυση βασίζεται στη γνήσια ποικιλία περιεχομένου.

3. Τα διπλότυπα μπορούν να γεμίσουν τη βάση δεδομένων και να δημιουργήσουν πλεονασμό στον χώρο αποθήκευσης. Με την κατάργησή τους, βελτιστοποιείται ο χώρος αποθήκευσης και βελτιώνεται η αποτελεσματικότητα της βάσης δεδομένων.
4. Καθώς μια βάση δεδομένων ειδήσεων μεγαλώνει με την πάροδο του χρόνου, το έργο της διαχείρισης και της διατήρησής της γίνεται πιο περίπλοκο. Η κατάργηση των διπλότυπων μειώνει τον φόρτο εργασίας όσον αφορά την εποπτεία και την επιμέλεια περιεχομένου.
5. Μια βάση δεδομένων που είναι γνωστή ότι περιέχει πολλά αντίγραφα ή σχεδόν αντίγραφα μπορεί να θεωρηθεί λιγότερο αξιόπιστη. Οι χρήστες ενδέχεται να αμφισβητήσουν την αξιοπιστία των πληροφοριών που παρουσιάζονται εάν συναντούν επανειλημμένα το ίδιο περιεχόμενο. Η διατήρηση μιας καθαρής και μοναδικής βάσης δεδομένων ενισχύει την αξιοπιστία της πλατφόρμας.
6. Σε ορισμένες περιπτώσεις, ενδέχεται να προκύψουν διπλότυπα άρθρα λόγω μη εξουσιοδοτημένης αντιγραφής ή λογοκλοπής. Καταργώντας τα διπλότυπα, είναι εφικτή η μείωση του κινδύνου παραβιάσεων πνευματικών δικαιωμάτων και νομικών ζητημάτων που σχετίζονται με την ιδιοκτησία περιεχομένου.

Συνοπτικά, η κατάργηση διπλότυπων ή σχεδόν διπλότυπων από μια βάση δεδομένων ειδήσεων είναι ζωτικής σημασίας για τη διατήρηση της ποιότητας των δεδομένων, τη βελτίωση της ακρίβειας της ανάλυσης, της εμπειρίας χρήστη και τη συμμόρφωση με τους κανονισμούς, μεταξύ άλλων πλεονεκτημάτων [AAE12], [HYK10].

Χρησιμοποιήθηκαν δύο βασικές μέθοδοι για την ανίχνευση και τη διαγραφή των διπλοτύπων άρθρων ειδήσεων μέσα στη βάση δεδομένων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

7.2.2 Πρώτη μέθοδος

Ακολουθεί η συνάρτηση **database_duplicates** που αφαιρεί τις διπλότυπες εγγραφές από έναν καθορισμένο πίνακα βάσης δεδομένων, διατηρώντας μόνο την καταχώρηση με το υψηλότερο αναγνωριστικό.

```
# Συνάρτηση για διαγραφή διπλοτύπων από το database
def database_duplicates(table_to_filter):
    # Καθορισμός της στήλης που περιέχει τους τίτλους
    title_column = "title"
    # Διαγραφή όλων των διπλοτύπων εκτός από αυτό με το υψηλότερο id
    sql = f"DELETE t1 FROM {table_to_filter} t1 INNER JOIN {table_to_filter} t2 WHERE t1.{title_column} = t2.{title_column} AND t1.id < t2.id"
    mycursor.execute(sql)
    mydb.commit()
```

Εικόνα 7-7: Συνάρτηση database_duplicates

Παρακάτω ακολουθεί η ανάλυση του παραπάνω κώδικα.

1. Ορίζεται μια συνάρτηση με το όνομα database_duplicates που παίρνει ένα όρισμα, **table_to_filter**, το οποίο είναι το όνομα του πίνακα της βάσης δεδομένων από τον οποίο θα αφαιρεθούν οι διπλές εγγραφές.
2. Ορίζεται η μεταβλητή **title_column** στη συμβολοσειρά "title". Προϋποθέτει ότι η στήλη τίτλου στον καθορισμένο πίνακα είναι αυτή που θα χρησιμοποιηθεί για τον εντοπισμό διπλών εγγραφών.
3. Εκτελεί την πραγματική κατάργηση διπλότυπων εγγραφών από τον καθορισμένο πίνακα. Κατασκευάζει ένα ερώτημα SQL που περιλαμβάνει μια λειτουργία αυτοσύνδεσης:
   - Το ερώτημα χρησιμοποιεί το όνομα του πίνακα που καθορίζεται από το **table_to_filter** και για τα ψευδώνυμα t1 και t2.
   - Η λειτουργία **INNER JOIN** αντιστοιχεί σε γραμμές όπου οι τίτλοι είναι ίδιοι (t1.{title_column} = t2.{title_column}) και όπου το αναγνωριστικό του t1 είναι μικρότερο από το αναγνωριστικό του t2 (t1.id < t2.id).
   - Η πρόταση **DELETE** καταργεί τις σειρές από τον πίνακα με βάση τις συνθήκες που καθορίζονται από το **INNER JOIN**.
   - Μετά την εκτέλεση του ερωτήματος διαγραφής χρησιμοποιώντας mycursor.execute(sql), οι αλλαγές δεσμεύονται στη βάση δεδομένων χρησιμοποιώντας mydb.commit().

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Συνοπτικά, η συνάρτηση **database_duplicates** αφαιρεί διπλότυπες εγγραφές από έναν καθορισμένο πίνακα με βάση τους αντίστοιχους τίτλους. Διατηρεί μόνο την καταχώρηση με το υψηλότερο αναγνωριστικό, εξαλείφοντας ουσιαστικά τις διπλές εγγραφές διατηρώντας παράλληλα την πιο πρόσφατη έκδοση.

7.2.3 Δεύτερη μέθοδος

Προτού αναλυθεί η δεύτερη μέθοδος, θα πρέπει να αναφερθεί η συνάρτηση **generate_new_id** όπου χρησιμοποιείται από αυτή. Η συνάρτηση **generate_new_id** υπολογίζει ένα νέο μοναδικό αναγνωριστικό για έναν καθορισμένο πίνακα βάσης δεδομένων.

```
1 def generate_new_id(mycursor, table_name):
2     # Παράγει ένα μοναδικό ID για τον duplicates table
3     mycursor.execute(f"SELECT MAX(id) FROM {table_name}")
4     max_id = mycursor.fetchone()[0]
5     new_id = max_id + 1 if max_id else 1
6     return new_id
```

Εικόνα 7-8: Συνάρτηση generate_new_id

1. Αυτή η γραμμή ορίζει μια συνάρτηση με το όνομα generate_new_id που παίρνει δύο ορίσματα:
   - **mycursor**, που είναι ένα αντικείμενο δρομέα για αλληλεπίδραση με τη βάση δεδομένων,
   - **table_name**, που είναι το όνομα του πίνακα βάσης δεδομένων για τον οποίο πρέπει να δημιουργηθεί ένα νέο μοναδικό αναγνωριστικό.
2. Δημιουργία νέου αναγνωριστικού:
   - Εκτελεί ένα ερώτημα SQL για να ανακτήσει τη μέγιστη (υψηλότερη) τιμή της στήλης **'id'** από τον καθορισμένο πίνακα. Η συνάρτηση **MAX(id)** ανακτά τη μέγιστη τιμή της στήλης **'id'**, η οποία θεωρείται ότι είναι πρωτεύον κλειδί που αυξάνεται αυτόματα.
   - Το αποτέλεσμα του ερωτήματος ανακτάται χρησιμοποιώντας τη **mycursor.fetchone()**, η οποία επιστρέφει μια πλειάδα. Το [0]

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

χρησιμοποιείται για την εξαγωγή της πρώτης (και μοναδικής) τιμής από την πλειάδα, η οποία είναι η μέγιστη τιμή αναγνωριστικού που βρέθηκε.

- Στη συνέχεια υπολογίζεται το νέο αναγνωριστικό που θα χρησιμοποιηθεί. Εάν το **max_id** έχει μια τιμή (που σημαίνει ότι υπάρχουν υπάρχουσες εγγραφές), το νέο αναγνωριστικό ορίζεται σε **max_id + 1**. Εάν το **max_id** είναι **None** (που σημαίνει ότι ο πίνακας είναι κενός), το νέο ID ορίζεται σε **1**.

3. Η συνάρτηση επιστρέφει την υπολογιζόμενη νέα τιμή ID.

Συνοπτικά, η συνάρτηση **generate_new_id** υπολογίζει ένα νέο μοναδικό αναγνωριστικό για έναν καθορισμένο πίνακα βάσης δεδομένων ζητώντας τη μέγιστη υπάρχουσα τιμή ID στον πίνακα και αυξάνοντάς την κατά 1. Αυτή η συνάρτηση είναι χρήσιμη για τη δημιουργία νέων αναγνωριστικών κατά την εισαγωγή εγγραφών σε έναν πίνακα με αυτόματη αυξημένο πρωτεύον κλειδί. Είναι ιδιαίτερα χρήσιμη όταν είναι επιθυμητή η εισαγωγή νέων εγγραφών σε έναν πίνακα, διασφαλίζοντας ταυτόχρονα ότι τα αναγνωριστικά είναι μοναδικά και ακολουθούν την καθιερωμένη σειρά. Υποβάλλοντας ερώτημα για το μέγιστο υπάρχον αναγνωριστικό στον πίνακα και δημιουργώντας ένα νέο αναγνωριστικό αυξάνοντάς το, η συνάρτηση εγγυάται ότι το νέο αναγνωριστικό δεν θα έρχεται σε διένεξη με κανένα υπάρχον αναγνωριστικό. Αυτή η προσέγγιση απλοποιεί τη διαδικασία εργασίας με πρωτεύοντα κλειδιά και διασφαλίζει την ακεραιότητα των δεδομένων κατά την εισαγωγή νέων δεδομένων στον πίνακα με αποτέλεσμα να τηρείτε η δομή του σχήματος της βάσης δεδομένων.

Στη δεύτερη μέθοδο, παρατίθεται η συνάρτηση **DB_duplicates** που προσδιορίζει και μετακινεί σχεδόν διπλότυπα άρθρα ειδήσεων από έναν καθορισμένο πίνακα βάσης δεδομένων σε έναν άλλο πίνακα που ονομάζεται **"duplicates"**. Αυτή η διαδικασία περιλαμβάνει σύγκριση των τίτλων των άρθρων και των ημερομηνιών δημοσίευσης.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

```python
def DB_duplicates(table_name):
    mycursor.execute(f"SELECT * FROM {table_name}")
    # Παίρνει όλες τις γραμμές από το παραπάνω query
    rows = mycursor.fetchall()
    df = pd.DataFrame(rows, columns=['id', 'title', 'category', 'description', 'link', 'content',
                                     'pubDate', 'sublevel', 'original_news_id'])

    # Καθορίζει το κατώφλι για την ομοιότητα μεταξύ τίτλων
    threshold = 0.55

    duplicates = []

    # Για κάθε ζευγάρι των νέων ειδήσεων
    for i in range(len(df)):
        for j in range(i + 1, len(df)):
            # Εξάγει την ημερομηνία και την ώρα δημοσίευσης από την στήλη pubDate
            date1 = df.iloc[i]['pubDate'][:19]  # εξαγωγή των πρώτων 19 χαρακτήρων (YYYY-MM-DD HH:MM:SS)
            date2 = df.iloc[j]['pubDate'][:19]  # εξαγωγή των πρώτων 19 χαρακτήρων (YYYY-MM-DD HH:MM:SS)
            # Υπολογίζει την ομοιότητα των τίτλων χρησιμοποιώντας SequenceMatcher library
            sim = SequenceMatcher(None, df.iloc[i]['title'],df.iloc[j]['title']).ratio()
            if sim >= threshold:
                # Εάν η ομοιότητα είναι πάνω από το όριο, γίνεται προσθήκη του ζεύγους στη λίστα διπλότυπων
                duplicates.append((df.iloc[i]['id'], df.iloc[j]['id'], max(date1, date2)))

    # Ταξινόμιση της λίστας των διπλότυπων κατά ημερομηνία δημοσίευσης με φθίνουσα σειρά
    duplicates.sort(key=lambda x: x[2], reverse=True)
    # print(duplicates)

    # Επανάληψη της λίστας των διπλότυπων και διαγραφή του νεότερου άρθρου ειδήσεων από τη βάση δεδομένων
    for d in duplicates:
        # Μετατροπή numpy.int64 σε regular int
        id_to_delete = int(d[1])
        # Ανάκτηση της σειρά που θα εισαχθεί στον πίνακα διπλότυπων
        mycursor.execute(f"SELECT * FROM {table_name} WHERE id = %s", (id_to_delete,))
        duplicate_row = mycursor.fetchone()
        if duplicate_row is not None:
            # Παραγωγή ενός νέου μοναδικού ID για την εισαγόμενη σειρά
            new_id = generate_new_id(mycursor, "duplicates")
            duplicate_row = list(duplicate_row)
            duplicate_row[0] = new_id
            duplicate_row = tuple(duplicate_row)
            # Εισαγωγή της εγγραφής στον πίνακα διπλότυπων
            insert_query = "INSERT INTO duplicates (id, title, category, description, link, content, pubDate, sublevel, original_news_id) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s)"
            mycursor.execute(insert_query, duplicate_row)
            mydb.commit()
        else:
            print("Error: could not retrieve duplicate row from database")
        # Διαγραφή της εγγραφής με το υψηλότερο pubDate (δηλαδή την πιο πρόσφατη) από τον αρχικό πίνακα
        mycursor.execute(f"DELETE FROM {table_name} WHERE id = %s", (id_to_delete,))
        mydb.commit()
        print("Moved news article with ID", d[1], "to the duplicates table")
```

Εικόνα 7-9: Συνάρτηση DB_duplicates

Παρακάτω γίνεται η ανάλυση της συγκεκριμένης συνάρτησης.

1. Ορίζεται η συνάρτηση με το όνομα **DB_duplicates** που παίρνει ένα όρισμα, **table_name**, το οποίο είναι το όνομα του πίνακα βάσης δεδομένων που περιέχει άρθρα ειδήσεων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2. Ο κώδικας ζητά από τον καθορισμένο πίνακα βάσης δεδομένων (table_name) να ανακτήσει όλες τις γραμμές. Στη συνέχεια, τα δεδομένα που λαμβάνονται χρησιμοποιούνται για τη δημιουργία ενός Dataframe (df) με στήλες που αντιστοιχούν στα διάφορα χαρακτηριστικά ενός άρθρου ειδήσεων.
3. Αυτή η γραμμή ορίζει μια τιμή κατωφλίου για την ομοιότητα μεταξύ των τίτλων των άρθρων. Άρθρα με ομοιότητα τίτλου πάνω από αυτό το όριο θεωρούνται πιθανά διπλότυπα. Η επιλογή ορίου ομοιότητας έγινε με πειραματικές μεθόδους προκειμένου να μην αποκόπτει σημαντικές ειδήσεις αλλά και να μην περιλαμβάνει όσο γίνεται διπλότυπες ειδήσεις.
4. Συγκρίνεται κάθε ζεύγος άρθρων ειδήσεων στο Dataframe χρησιμοποιώντας τους τίτλους τους και υπολογίζει την ομοιότητα χρησιμοποιώντας το SequenceMatcher από τη βιβλιοθήκη difflib. Εάν η ομοιότητα μεταξύ των τίτλων είναι πάνω από το όριο, τα αναγνωριστικά των άρθρων και η υψηλότερη από τις δύο ημερομηνίες δημοσίευσης προστίθενται στη λίστα διπλότυπων.
5. Η λίστα διπλότυπων ταξινομείται με φθίνουσα σειρά με βάση τις ημερομηνίες δημοσίευσης των πιθανών διπλότυπων ζευγών.
6. Για κάθε διπλότυπο ζεύγος στην ταξινομημένη λίστα, ο κώδικας κάνει τα εξής:
    - Ανακτά τις λεπτομέρειες του διπλότυπου άρθρου από τον αρχικό πίνακα.
    - Δημιουργεί ένα νέο μοναδικό αναγνωριστικό για τη διπλότυπη καταχώρηση.
    - Εισάγει τις λεπτομέρειες του διπλότυπου άρθρου στον πίνακα "διπλότυπα".
    - Διαγράφει το διπλότυπο άρθρο από τον αρχικό πίνακα.

Συνοπτικά, η συνάρτηση **DB_duplicates** προσδιορίζει σχεδόν διπλότυπα άρθρα ειδήσεων σε έναν καθορισμένο πίνακα βάσης δεδομένων, μετακινεί αυτά τα αντίγραφα στον πίνακα «duplicates» με ενημερωμένα αναγνωριστικά και τα αφαιρεί από τον αρχικό πίνακα. Αυτή η διαδικασία περιλαμβάνει τη σύγκριση των τίτλων των άρθρων για ομοιότητα και τον χειρισμό των διπλότυπων με βάση τις ημερομηνίες δημοσίευσής τους.

7.2.4 Επίλογος

Συμπερασματικά, αναλύθηκαν και επεξηγήθηκαν τρεις αξιόλογες λειτουργίες που φωτίζουν την τέχνη του αποτελεσματικού χειρισμού της βάσης δεδομένων. Η συνάρτηση **database_duplicates** έδειξε την ικανότητά της εντοπίζοντας και

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

εξαλείφοντας διπλότυπες καταχωρήσεις, διασφαλίζοντας την ακεραιότητα των δεδομένων διατηρώντας τις πιο πρόσφατες εκδόσεις. Η συνάρτηση **generate_new_id** χειρίστηκε με χάρη το έργο της δημιουργίας μοναδικών αναγνωριστικών για νέες εγγραφές, διατηρώντας την ακολουθία των πρωτευόντων κλειδιών με λεπτότητα. Τέλος, η συνάρτηση **DB_duplicates** είχε την ικανότητα ανίχνευσης παρόμοιων ειδήσεων, μετακινώντας τα αντίγραφα σε ένα καθορισμένο πίνακα.

Αυτές οι λειτουργίες συλλογικά πραγματοποιήθηκαν χρησιμοποιώντας τις βιβλιοθήκες και τις λειτουργίες της Python, όπου συνέβαλαν στην ενοποίηση της βάσης δεδομένων, του χειρισμού SQL και της ανάλυσης δεδομένων. Αυτό αποτελεί απόδειξη της ικανότητας του προγραμματισμού για απρόσκοπτη πλοήγηση στην πολυπλοκότητα των βάσεων δεδομένων, ενισχύοντας ισχυρές εφαρμογές και γνώσεις που βασίζονται σε δεδομένα. Συμπερασματικά, γίνεται αντιληπτό το βαθύ δυναμικό που προκύπτει από τη συνέργεια μεταξύ προγραμματισμού και διαχείρισης βάσεων δεδομένων, δίνοντας τη δυνατότητα για κυριαρχία των δεδομένα τόσο με ακρίβεια όσο και με δημιουργικότητα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

8 Εντοπισμός ειδησεογραφικών άρθρων που περιγράφουν νεότερες εξελίξεις για γνωστά γεγονότα

8.1 Εισαγωγή

Οι επώνυμες οντότητες χρησιμεύουν ως κρίσιμα στοιχεία στη διαδικασία ανάλυσης άρθρων ειδήσεων που περιγράφουν νεότερες εξελίξεις για γνωστά γεγονότα, γνωστά ως follow-up. Αυτές οι οντότητες περιλαμβάνουν συγκεκριμένες λέξεις ή φράσεις που αναφέρονται σε μοναδικά άτομα, οργανισμούς, τοποθεσίες, ημερομηνίες και άλλα σχετικά στοιχεία. Οι επώνυμες οντότητες συχνά περιλαμβάνουν βασικούς χαρακτήρες, τοποθεσίες και άλλες σημαντικές πτυχές μιας ιστορίας. Στον τομέα της ανάλυσης ειδησεογραφικών άρθρων, η χρήση επώνυμων οντοτήτων προσφέρει ευδιάκριτα πλεονεκτήματα που την ξεχωρίζουν από εναλλακτικές μεθόδους κατά την αναζήτηση επακόλουθων πληροφοριών και γνώσεων.

Ένα πρωταρχικό πλεονέκτημα των επώνυμων οντοτήτων είναι η εγγενής τους ικανότητα να διευκολύνουν την ακρίβεια στην ανάκτηση πληροφοριών. Όταν πραγματοποιείται αναζήτηση νέων άρθρων, αξιοποιούνται οι επώνυμες οντότητες για να εντοπιστούν ακριβή άτομα, οργανισμοί ή τοποθεσίες που σχετίζονται με την αρχική ιστορία. Αυτή η εστιασμένη προσέγγιση μειώνει σημαντικά την πιθανότητα απώλειας σχετικών ενημερώσεων, διασφαλίζοντας ότι τα επόμενα άρθρα σχετίζονται άμεσα με τα βασικά στοιχεία ενδιαφέροντος.

Οι επώνυμες οντότητες βελτιστοποιούν επίσης τη διαδικασία παρακολούθησης της εξέλιξης των γεγονότων με την πάροδο του χρόνου. Συγκεκριμένες οντότητες, όπως ημερομηνίες, ώρες και τοποθεσίες επιτρέπουν τη χρονολογική οργάνωση των άρθρων, επιτρέποντας τη δημιουργία ενός συνεκτικού χρονοδιαγράμματος των εξελίξεων. Αυτό το χρονολογικό πλαίσιο ενισχύει την ανάλυση του πώς έχουν εξελιχθεί τα γεγονότα και πώς διάφορες οντότητες έχουν επηρεαστεί ή εμπλακεί σε διαφορετικά στάδια.

Σε αντίθεση με τις εναλλακτικές μεθόδους, όπως οι αναζητήσεις που βασίζονται σε λέξεις-κλειδιά, οι επώνυμες οντότητες προσφέρουν μια πιο λεπτή κατανόηση των σχέσεων μεταξύ των διαφόρων στοιχείων μιας είδησης. Οι επώνυμες οντότητες δημιουργούν εγγενώς συνδέσεις μεταξύ ατόμων, οργανισμών και γεγονότων, επιτρέποντας την εξερεύνηση της περίπλοκης αλληλεπίδρασης και των εξαρτήσεων που προκύπτουν κατά την εξέλιξη ενός γεγονότος.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

8.2 Εντοπισμός ειδησεογραφικών άρθρων εξέλιξης γεγονότος

Μία από τις απλούστερες και πιο κοινές στρατηγικές που χρησιμοποιούνται στην αναγνώριση επώνυμης οντότητας (named entity recognition NER) είναι η χρήση κεφαλαίων. Η χρήση κεφαλαίων στο πρώτο γράμμα μιας λέξης είναι μια ισχυρή ένδειξη ότι η λέξη μπορεί να είναι μια επώνυμη οντότητα, ιδιαίτερα ένα κύριο ουσιαστικό. Τα σωστά ουσιαστικά αναφέρονται συχνά σε συγκεκριμένες οντότητες όπως άτομα, οργανισμούς ή τοποθεσίες.

Συγκεκριμένα, η χρήση κεφαλαίων βοηθά στην αναγνώριση επώνυμων οντοτήτων:

1. Αρχικό γράμμα κεφαλαίο: Πολλές επώνυμες οντότητες, όπως ονόματα ατόμων, ονόματα εταιρειών, πόλεων και χωρών, ξεκινούν με κεφαλαίο γράμμα, οπότε αντιμετωπίζονται ως πιθανές ονομαστικές οντότητες.
2. Διαφοροποίηση ίδιων λέξεων: Η χρήση κεφαλαίων θέτει επώνυμες οντότητες εκτός από τα κοινά ουσιαστικά. Για παράδειγμα, ενώ το "apple" είναι ένα κοινό ουσιαστικό, το "Apple" που αναφέρεται στην εταιρεία τεχνολογίας είναι μια ονομαστική οντότητα. Το αρχικό κεφαλαίο γράμμα βοηθά να διαφοροποιήθουν τα δύο.
3. Συμφραζόμενα: Ενώ η χρήση κεφαλαίων είναι ισχυρός δείκτης, το πλαίσιο είναι επίσης σημαντικό. Για παράδειγμα, στην πρόταση "He ate an apple", η λέξη "apple" δεν είναι μια επώνυμη οντότητα. Ωστόσο, στην πρόταση "Εργάζεται στην Apple", το κεφαλαίο "Apple" πιθανότατα αναφέρεται στην εταιρεία και είναι μια επώνυμη οντότητα.

Ενώ η χρήση κεφαλαίων είναι μια χρήσιμη δυνατότητα, είναι σημαντικό να σημειωθεί ότι δεν είναι όλες οι λέξεις με κεφαλαία επώνυμες οντότητες. Τα ακρωνύμια, οι τίτλοι και ορισμένες άλλες λέξεις μπορεί επίσης να γράφονται με κεφαλαία, αλλά δεν αντιπροσωπεύουν απαραίτητα ονομαστικές οντότητες.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

8.2.1 Συνάρτηση Find_words

Πριν αναλύσουμε τη συγκεκριμένη συνάρτηση, πρέπει να αναφερθούμε στη δημιουργία του λεξικού **capitalized_words.**

```
# Χρήση regex για εύρεση λέξεων με κεφαλαία σε κάθε τίτλο, εξαιρουμένων εκείνων που ακολουθούν μια τελεία, και αποθηκεύση τους σε ένα λεξικό με το id ως τιμή
capitalized_words = {}
for index, row in df.iterrows():
    title_id = row['id']
    title = row['title']
    words = re.findall(r'(?<!\.)\b(?!' + '|'.join(capital_words) + r')[A-Ω]\w*', title, flags=re.U)
    for word in words:
        if word not in capitalized_words:
            capitalized_words[word] = [title_id]
        else:
            capitalized_words[word].append(title_id)
```

Εικόνα 8-1 Δημιουργία λεξικού capitalized_words

Αυτός ο κώδικας χρησιμοποιεί το Regular Expression (regex) για να εντοπίσει λέξεις που περιέχουν κεφαλαία γράμματα σε τίτλους και περιγραφές ενός DataFrame.

1. `capitalized_words = {}`: Δημιουργεί ένα κενό λεξικό με το όνομα `capitalized_words` που θα χρησιμοποιηθεί για να αποθηκεύσει τα αποτελέσματα της εύρεσης λέξεων με κεφαλαία γράμματα.
2. βρόχος `for index, row in df.iterrows():` για τους τίτλους (`title`) του DataFrame.
3. Στην επανάληψη κάθε τίτλου, αποθηκεύεται το `id` του τίτλου ή της περιγραφής στη μεταβλητή `title_id`.
4. Η εκάστοτε σειρά του DataFrame αποθηκεύεται στις μεταβλητές `title`.
5. Η συνάρτηση `re.findall` χρησιμοποιείται για να εντοπίσει όλες τις λέξεις που περιέχουν κεφαλαία γράμματα στον τίτλο. Ο παρακάτω regex pattern εφαρμόζεται:
    a. `(?<!\.)`: Αυτή η έκφραση δηλώνει ότι η λέξη δεν πρέπει να ακολουθείται από τελεία.
    b. `\b`: Ταιριάζει το όριο μιας λέξης.
    c. `(?!' + '|'.join(capital_words) + r')`: Αυτό το τμήμα αποτρέπει την επανάληψη ενός ίδιου κεφαλαίου λέξης.
    d. `[A-Ω]\w*`: Ταιριάζει μια λέξη που αρχίζει με κεφαλαίο γράμμα από το ελληνικό αλφάβητο, ακολουθούμενο από μη κεφαλαία γράμματα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

6. Οι λέξεις που εντοπίζονται προστίθενται στο λεξικό `capitalized_words`, με το `title_id` ως τιμή. Εάν μια λέξη δεν υπάρχει στο λεξικό, δημιουργείται μια εισαγραφή με τη λέξη ως κλειδί και μια λίστα που περιέχει το `title_id`. Εάν η λέξη υπάρχει ήδη, το `title_id` προστίθεται στη λίστα που αντιστοιχεί στη λέξη.

Συνοπτικά, αυτός ο κώδικας εντοπίζει λέξεις με κεφαλαία γράμματα σε τίτλους, αποφεύγοντας τις περιπτώσεις που ακολουθούν μια τελεία, και τις αποθηκεύει σε ένα λεξικό με το `id` ως τιμή. Ακριβώς η ίδια διαδικασία ακολουθείται και για την περιγραφή της κάθε είδησης.

Η συνάρτηση **find_words** αναζητά έναν πίνακα βάσης δεδομένων για άρθρα ειδήσεων που περιέχουν συγκεκριμένες λέξεις από το δεδομένο λεξικό **capitalized_words**. Επεξεργάζεται αυτά τα άρθρα και επιστρέφει μια λίστα ζευγών, καθένα από τα οποία αποτελείται από αναγνωριστικά που σχετίζονται με τα αντίστοιχα άρθρα ειδήσεων και λέξεις λεξικού. Φαίνεται παρακάτω ο κώδικας της συνάρτησης:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

```python
def find_words(table):
    display(f"\n {table}")
    # Δημιουργήστε μια κενή λίστα για να αποθηκεύσετε τα αποτελέσματα που ταιριάζουν
    matching_results = []
    # Για κάθε λέξη στο λεξικό με κεφαλαία_λέξεις, αναζητήστε τη στη βάση δεδομένων
    for word, ids in capitalized_words.items():
        for id in ids:
            # Αναζήτηση της πρώτης καταχώρισης ειδήσεων που περιέχει τη λέξη στον τίτλο ή την περιγραφή
            query = f"SELECT id, title FROM {table} WHERE title OR description LIKE '%{word}%' LIMIT 1"
            mycursor.execute(query)
            result = mycursor.fetchone()

            if result is not None:

                # Το σύνολο των λέξεων που ταιριάζουν στον τίτλο
                title_words = set(w for w in result[1].split() if w in capitalized_words)

                # Το σύνολο των λέξεων που ταιριάζουν στο λεξικό capitalized_words
                dict_words = set(word for word in title_words if word in capitalized_words)
                # Τομή των συνόλων
                common_words = title_words.intersection(dict_words)

                if len(common_words) >= 2:
                    # Προσθήκη πλειάδας που περιέχει τα δύο αναγνωριστικά στη λίστα
                    matching_results.append((int(id), int(result[0])))

    # Επιστροφή της λίστας των αποτελεσμάτων που ταιριάζουν
    return matching_results
```

Εικόνα 8-2: Συνάρτηση find_words

Η συνάρτηση αυτή καλείται ως **crime_accident_list = find_words("crime_accident")** στον κώδικα. Ακολουθεί η ανάλυση του τι κάνει ο κώδικας:

1. Ο κώδικας ορίζει μια συνάρτηση Python με το όνομα **find_words** που παίρνει ένα όρισμα που ονομάζεται πίνακας. Αυτό το όρισμα θα πρέπει να είναι το όνομα ενός πίνακα βάσης δεδομένων όπου η συνάρτηση θα αναζητήσει αντίστοιχα άρθρα ειδήσεων.
2. Η συνάρτηση ξεκινά εμφανίζοντας το όνομα του πίνακα βάσης δεδομένων που επεξεργάζεται αυτήν τη στιγμή χρησιμοποιώντας τη λειτουργία εμφάνισης. Αυτή η έξοδος προορίζεται για να δείξει ποιος πίνακας είναι υπό επεξεργασία.
3. Η συνάρτηση προετοιμάζει μια κενή λίστα με το όνομα **matching_results**. Αυτή η λίστα θα χρησιμοποιηθεί για την αποθήκευση ζευγαριών

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αναγνωριστικών που σχετίζονται με αντιστοιχισμένα άρθρα ειδήσεων και λέξεις λεξικού.

4. Η συνάρτηση επαναλαμβάνεται μέσω των στοιχείων ενός λεξικού που ονομάζεται **capitalized_words**. Αυτό το λεξικό περιέχει λέξεις με κεφαλαία ως κλειδιά και αντίστοιχα αναγνωριστικά ως τιμές. Για κάθε λέξη στο λεξικό και για κάθε αναγνωριστικό που σχετίζεται με αυτήν τη λέξη, η συνάρτηση δημιουργεί ένα ερώτημα SQL για αναζήτηση στον παρεχόμενο πίνακα για άρθρα ειδήσεων όπου ο τίτλος ή η περιγραφή περιέχει την τρέχουσα λέξη. Το ερώτημα εκτελείται χρησιμοποιώντας το αντικείμενο του δρομέα "**mycursor**". Εάν βρεθεί ένα αποτέλεσμα, η συνάρτηση εξάγει σχετικές πληροφορίες από το αποτέλεσμα και προχωρά στην επεξεργασία τους.
5. Η συνάρτηση ελέγχει εάν βρέθηκε ένα αποτέλεσμα στο ερώτημα της βάσης δεδομένων. Εάν υπάρχει αποτέλεσμα, επεξεργάζεται τον τίτλο του άρθρου για να βρει λέξεις που ταιριάζουν τόσο με το λεξικό όσο και με τον τίτλο του άρθρου. Δημιουργεί σύνολα λέξεων από τον τίτλο και το λεξικό και υπολογίζει την τομή τους. Εάν η τομή περιέχει τουλάχιστον δύο λέξεις, η συνάρτηση προσθέτει μια πλειάδα που περιέχει το αναγνωριστικό του άρθρου και το αναγνωριστικό του αποτελέσματος στη λίστα **matching_results**.
6. Η συνάρτηση επιστρέφει τη λίστα **matching_results**, η οποία περιέχει πλέον ζεύγη αναγνωριστικών που σχετίζονται με άρθρα ειδήσεων και λέξεις λεξικού που ταιριάζουν.

Συνοπτικά, η συνάρτηση **find_words** αναζητά συγκεκριμένες λέξεις από ένα λεξικό σε έναν πίνακα βάσης δεδομένων, προσδιορίζει άρθρα ειδήσεων που περιέχουν αυτές τις λέξεις και επιστρέφει ζεύγη αναγνωριστικών που συνδέουν τα αντίστοιχα άρθρα και λέξεις λεξικού. Αυτή η λειτουργία χρησιμοποιείται για εργασίες όπως ο εντοπισμός άρθρων ειδήσεων που σχετίζονται με συγκεκριμένες λέξεις-κλειδιά.

Αυτή η εντολή χρησιμοποιείται και για τις υπόλοιπες 10 κατηγορίες γεγονότων και η μόνη αλλαγή που γίνεται είναι πως στην κλήση της συνάρτησης αλλάζει ο πίνακας της κατηγορίας.

8.2.2 Συνάρτηση Follow_up

Η συνάρτηση αυτή χρησιμοποιείται για την ενημέρωση ενός Dataframe χρησιμοποιώντας τις πληροφορίες από μια λίστα ζευγών προκειμένου να βρεθεί η εξέλιξη των γεγονότων όπου αυτό είναι εφικτό. Παρατίθεται ο κώδικας παρακάτω:

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

```
def follow_up(df, list):
    for entry in list:
        article_id = entry[0]
        found_article_id = entry[1]

        # Ενημέρωση της επιθυμητής σειράς στο dataframe χρησιμοποιώντας τη συνάρτηση .loc
        df.loc[df['id'] == article_id, 'original_news_id'] = found_article_id
```

Εικόνα 8-3: Συνάρτηση follow_up

Η συνάρτηση αυτή καλείται ως **follow_up(total_crime_accident_df, crime_accident_list)** στον κώδικα. Ακολουθεί η ανάλυση του τι κάνει ο κώδικας:

1. Αυτή η γραμμή καλεί τη συνάρτηση **follow_up** και της μεταβιβάζει δύο ορίσματα: **total_crime_accident_df** και **crime_accident_list**. Το πρώτο όρισμα, **total_crime_accident_df**, θεωρείται ότι είναι ένα Dataframe που περιέχει πληροφορίες σχετικά με άρθρα εγκλημάτων και ατυχημάτων. Το δεύτερο όρισμα, **crime_accident_list**, θεωρείται ότι είναι μια λίστα ζευγών.
2. Αυτή η γραμμή ορίζει μια συνάρτηση με το όνομα follow_up που παίρνει δύο ορίσματα: ένα Dataframe (df) και μια λίστα (list). Ο σκοπός της συνάρτησης είναι να ενημερώσει το Dataframe με βάση τις πληροφορίες που παρέχονται στη λίστα.
3. Μέσα στη συνάρτηση follow_up, ο κώδικας επαναλαμβάνεται σε κάθε καταχώρηση στην παρεχόμενη λίστα. Κάθε καταχώρηση θεωρείται ότι είναι ένα ζευγάρι αναγνωριστικών άρθρων: article_id και found_article_id. Αυτά τα αναγνωριστικά χρησιμοποιούνται για τη σύνδεση σχετικών άρθρων στο Dataframe.
4. Εντός του βρόχου, ο κώδικας χρησιμοποιεί τη συνάρτηση .loc του Dataframe για να εντοπίσει συγκεκριμένες σειρές όπου η στήλη "id" ταιριάζει με την τιμή του article_id. Στη συνέχεια, ενημερώνει την αντίστοιχη καταχώρηση στη στήλη 'original_news_id' με την τιμή found_article_id. Αυτή η λειτουργία δημιουργεί μια σύνδεση μεταξύ ενός άρθρου και του σχετικού αρχικού άρθρου ειδήσεων, ενημερώνοντας το αναγνωριστικό του αρχικού άρθρου ειδήσεων στο Dataframe.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

5. Η στήλη ‘original_news_id’ εισάγεται στο Dataframe της κάθε κατηγορίας στην συνάρτηση categorization που αναλύεται στην Ενότητα (8.13) της εργασίας [Σ23], που πραγματοποιεί την κατηγοριοποίηση των ειδησεογραφικών δεδομένων.

   Η εντολή μέσα σε αυτή τη συνάρτηση είναι η: **"total_crime_accident_df.insert(loc=7, column='original_news_id', value='0')":** Αυτή η γραμμή εισάγει μια άλλη νέα στήλη με το όνομα "original_news_id" στη θέση 7 στο DataFrame "total_crime_accident_df". Η παράμετρος «loc» καθορίζει τη θέση για εισαγωγή, η παράμετρος «column» καθορίζει το όνομα της νέας στήλης («original_news_id») και η παράμετρος «value» καθορίζει την τιμή '0' ώστε να εκχωρηθεί σε όλες τις εγγραφές στη νέα στήλη. Η στήλη αυτή χρησιμοποιείται ώστε όταν προκύψει μια είδηση που είναι εξέλιξη κάποια άλλης τότε στην στήλη αυτή τοποθετείται το id της αρχικής-μητρικής είδησης προκειμένου να επισημανθεί ως follow up είδηση. Αυτή η γραμμή ουσιαστικά προσθέτει μια νέα στήλη «original_news_id» στο DataFrame και τη συμπληρώνει με την τιμή '0' για όλες τις σειρές.

Συνοπτικά, η συνάρτηση **follow_up** παίρνει ως ορίσματα ένα Dataframe και μια λίστα ζευγών. Επαναλαμβάνεται στη λίστα και για κάθε ζευγάρι, ενημερώνει το αρχικό αναγνωριστικό του άρθρου ειδήσεων στο Dataframe για να δημιουργήσει μια σύνδεση μεταξύ σχετικών άρθρων. Ο κώδικας έχει σχεδιαστεί για να βελτιώνει τις πληροφορίες του Dataframe με βάση την παρεχόμενη λίστα ζευγών, δηλώνοντας έτσι την εξέλιξη του κάθε γεγονότος.

Αυτή η συνάρτηση χρησιμοποιείται και για τις υπόλοιπες 10 κατηγορίες γεγονότων και η μόνη αλλαγή που γίνεται είναι πως αλλάζει ο πίνακας της κατηγορίας ως όρισμα στην κλήση της συνάρτησης, καθώς και η αντίστοιχη λίστα της κατηγορίας.

8.2.3 Χρονικό πλαίσιο για την εύρεση follow up ειδήσεων

Διαπιστώθηκε πως η συνάρτηση Follow_up που αναλύθηκε στην ενότητα 2.6.2.2 δεν αρκεί. Αναλυτικότερα, λόγω του μεγάλου όγκου των ειδήσεων που προκύπτουν, δεν θα ήταν αποδοτικός ο κώδικας αν για κάθε είδηση εξέταζε αν είναι εξέλιξης κάποιας άλλης. Αυτό θα είχε ως αποτέλεσμα την μη σωστή διαχείριση του συστήματος όσον αφορά τους πόρους αλλά και φυσικά την αποδοτικότητά του. Για τον λόγο αυτό πάρθηκε η απόφαση να εξετάζει ο κώδικας για κάθε νέα είδηση, τρεις μέρες πριν για να βρει τυχόν την αρχική είδηση ή την εξέλιξη της παρούσας είδησης.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Αυτό επιτυγχάνεται με την ενημέρωση της συνάρτησης της παραπάνω υποενότητας όπως φαίνεται παρακάτω.

```
def follow_up(df, list, table_name):
    three_days_ago = datetime.now() - timedelta(days=3)

    for entry in list:
        article_id = entry[0]
        found_article_id = entry[1]

        # Μορφή της μεταβλητής three_days_ago σαν συμβολοσειρά σε μορφή 'YYYY-MM-DD HH:MM:SS'
        formatted_three_days_ago = three_days_ago.strftime('%Y-%m-%d %H:%M:%S')
        # print("formatted_three_days_ago: ", formatted_three_days_ago)

        # Προσθήκη ενός φίλτρου ημερομηνίας μέσω SQL query για την αναζήτηση follow-up ειδήσεων τις τελευταίες 3 μέρες
        query = f"SELECT id, title FROM {table_name} WHERE id = {found_article_id} AND pubdate >= '{formatted_three_days_ago}'"
        mycursor.execute(query)
        result = mycursor.fetchone()
        # print("result: ", result)

        if result is not None:
            # Ενημέρωση της επιθυμητής σειράς στο dataframe χρησιμοποιώντας τη συνάρτηση .loc
            df.loc[df['id'] == article_id, 'original_news_id'] = found_article_id
```

Εικόνα 8-4: Συνάρτηση follow up με χρονικό πλαίσιο

Το παραπάνω block κώδικα εκτελεί την εύρεση follow up ειδήσεων μέχρι και τρεις μέρες πριν. Συγκεκριμένα:

- Γίνεται λήψη της τρέχουσας ημερομηνίας και ώρας υπολογίζοντας ένα αντικείμενο ημερομηνίας που αντιπροσωπεύει πριν από τρεις ημέρες από την τρέχουσα ημερομηνία και ώρα χρησιμοποιώντας τη συνάρτηση **datetime.now()** και αφαιρώντας ένα **timedelta** τριών ημερών.
- Ο κώδικας επαναλαμβάνεται μέσω μιας λίστας που ονομάζεται λίστα, η οποία φαίνεται να περιέχει ζεύγη αναγνωριστικών άρθρων. Για κάθε ζεύγος στη λίστα, εξάγει το **άρθρο_id** και **found_article_id**.
- Μορφοποίηση της ημερομηνίας μετατρέποντας το αντικείμενο ημερομηνίας πριν από τρεις ημέρες σε μια μορφοποιημένη συμβολοσειρά στη μορφή «**YYYY-MM-DD HH:MM:SS**».
- Κατασκευάζεται μια συμβολοσειρά ερωτήματος SQL για την επιλογή δεδομένων από έναν πίνακα με το όνομα table_name όπου το αναγνωριστικό

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αντιστοιχεί στο **found_article_id** και η ημερομηνία έκδοσης είναι μεγαλύτερη ή ίση με τη μορφοποιημένη χρονική σήμανση πριν από τρεις ημέρες. Στη συνέχεια, εκτελείται το ερώτημα SQL και ανακτά μία σειρά αποτελεσμάτων.

- Εάν βρεθεί ένα αποτέλεσμα (δηλαδή, το αποτέλεσμα δεν είναι None), ενημερώνει το DataFrame (df) χρησιμοποιώντας τη μέθοδο **.loc**. Τροποποιεί τη στήλη **'original_news_id'** στο DataFrame όπου το **'id'** αντιστοιχεί στο **article_id**, ορίζοντας το σε **found_article_id**.

Συνολικά, αυτός ο κώδικας έχει σχεδιαστεί για να βρίσκει και να ενημερώνει πληροφορίες σε ένα DataFrame, με βάση δεδομένα που ανακτώνται από έναν πίνακα βάσης δεδομένων χρησιμοποιώντας ερωτήματα SQL. Η συνάρτηση χρησιμοποιείται για τη σύνδεση των πιο πρόσφατων γεγονότων με τις ειδήσεις που βρίσκονται στη βάση ανάλογα την ημερομηνία δημοσίευσης και τα αναγνωριστικά τους.

Αξίζει να σημειωθεί πως η διαδικασία αυτή αποφάνθηκε χρονοβόρα λόγω του μεγάλου όγκου των υπολογισμών που πρέπει να γίνουν στο μεγάλο πλήθος ειδησεογραφικών γεγονότων που εξετάζει και μετατρέπει. Είναι απαραίτητη η αναθεώρηση και βελτίωση του κώδικα σε αυτό το συγκεκριμένο τμήμα προκειμένου να επιτευχθεί μια πιο αποδοτική λειτουργία του συστήματος, χωρίς περιττές καθυστερήσεις.

8.3 Επίλογος

Συμπερασματικά, οι επώνυμες οντότητες παρουσιάζουν μια συναρπαστική και πλεονεκτική προσέγγιση για την ανάλυση επακόλουθων ειδήσεων. Η ακρίβειά τους, η φύση που δείχνει τη σημασία, το χρονολογικό τους πλαίσιο και η ικανότητα αποκάλυψης σχέσεων καθιστούν τις επώνυμες οντότητες μια ανώτερη επιλογή για συστηματική και ολοκληρωμένη αποκάλυψη των εξελισσόμενων στρωμάτων αφηγήσεων ειδήσεων. Στην πράξη, αυτό επιτυγχάνεται με τις συναρτήσεις "find_words" και "follow_up", οι οποίες έχουν συμβάλει στην εξερεύνηση της αλληλεπίδρασης βάσεων δεδομένων που βασίζεται σε Python και του χειρισμού δεδομένων, προσφέροντας πρακτικές και διορατικές δυνατότητες.

Η συνάρτηση "find_words" εμφανίζει την εφαρμογή της επεξεργασίας φυσικής γλώσσας σε ένα πλαίσιο βάσης δεδομένων. Διευκολύνοντας την αναζήτηση συγκεκριμένων λέξεων ή φράσεων μέσα στο κείμενο, αυτή η λειτουργία δίνει τη δυνατότητα στους χρήστες να αποκαλύπτουν σχετικές πληροφορίες. Όχι μόνο ενισχύει

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

την ικανότητα της ανάκτησης στοχευμένων δεδομένων, αλλά ρίχνει φως σε μοτίβα κειμένου που διαφορετικά θα μπορούσαν να παραμείνουν κρυφά. Αυτή η λειτουργία αποδεικνύεται ιδιαίτερα πολύτιμη όταν υπάρχει μεγάλος όγκος πληροφοριών κειμένου, επιτρέποντας στους χρήστες να εξάγουν αποτελεσματικά ουσιαστικό περιεχόμενο.

Από την άλλη πλευρά, η συνάρτηση "follow_up" αντιμετωπίζει τη δυναμική φύση των ενημερώσεων της βάσης δεδομένων. Επιτρέπει τη συστηματική προσαρμογή των καταχωρίσεων δεδομένων με βάση μια παρεχόμενη λίστα ζευγών. Αυτή η λειτουργία είναι ιδιαίτερα χρήσιμη σε σενάρια όπου οι σχέσεις δεδομένων εξελίσσονται με την πάροδο του χρόνου. Με την απρόσκοπτη ενημέρωση των εγγραφών σύμφωνα με τα σχετικά αναγνωριστικά, διατηρεί την ακεραιότητα και τη συνοχή της βάσης δεδομένων. Η λειτουργία "follow_up" ευθυγραμμίζεται με σενάρια του πραγματικού κόσμου όπου ο συγχρονισμός και η συντήρηση δεδομένων είναι ζωτικής σημασίας για τη διασφάλιση ακριβών και ενημερωμένων πληροφοριών.

Ουσιαστικά, μέσα από αυτές τις λειτουργίες τονίζονται οι πραγματικές δυνατότητες της Python στη διαχείριση βάσεων δεδομένων και την εξαγωγή ουσιαστικών πληροφοριών. Αξιοποιώντας τις δυνατότητες των "find_words" και "follow_up", ξεκλειδώνουμε ευκαιρίες για αποτελεσματική πλοήγηση σε δεδομένα κειμένου και διευκολύνουμε τις συνεχείς ενημερώσεις δεδομένων. Αυτές οι λειτουργίες αποτελούν παράδειγμα της συγχώνευσης τεχνολογίας και πρακτικότητας, υπογραμμίζοντας τη βαθιά επίδραση της Python στη βελτίωση των εργασιών που σχετίζονται με δεδομένα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

9 Πειραματική μελέτη

9.1 Εισαγωγή

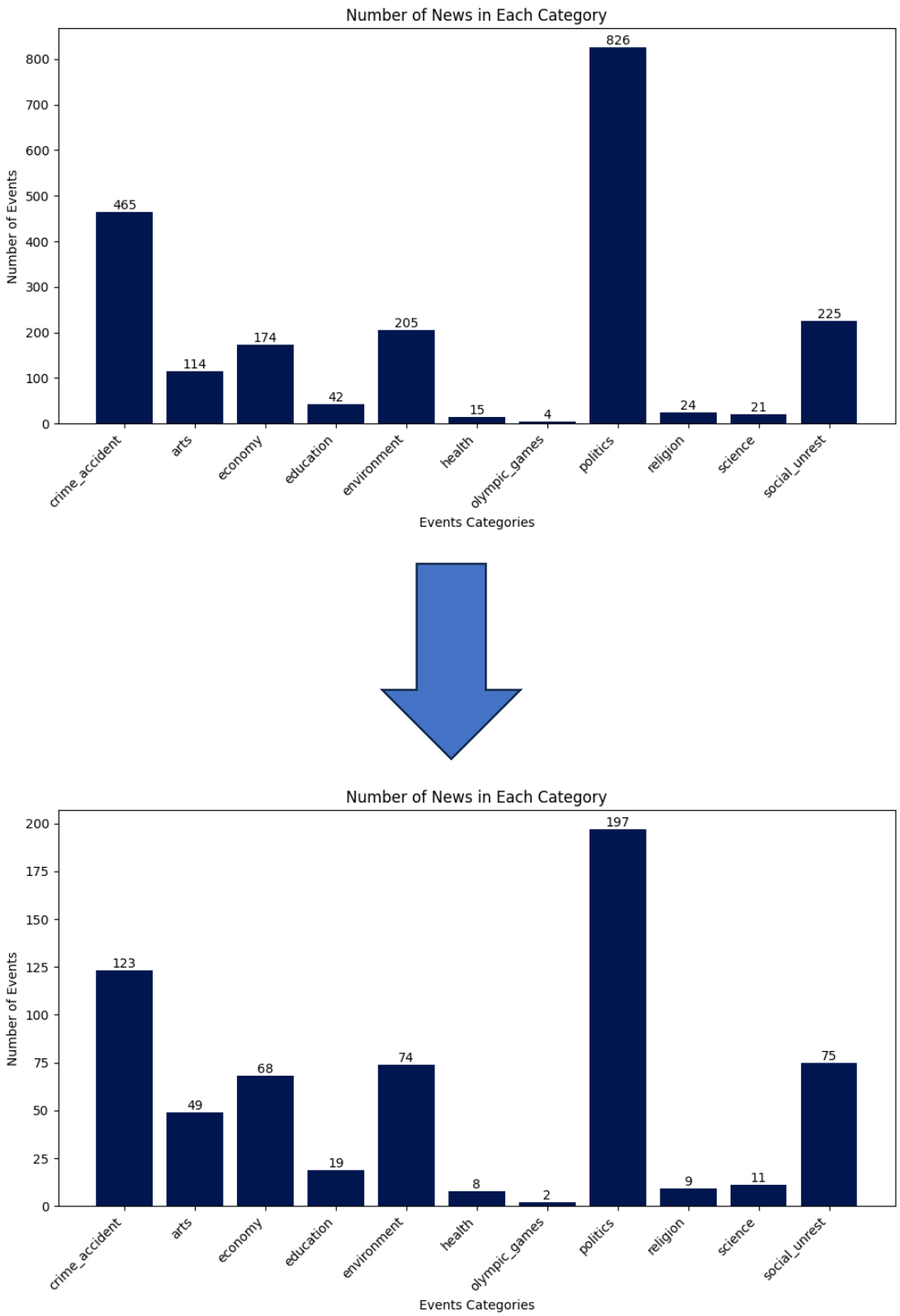
Για τη διεξαγωγή της πειραματικής μελέτης και, ως αποτέλεσμα, την αξιολόγηση της λειτουργικότητας του κώδικα της συγκεκριμένης πτυχιακής εργασίας, πραγματοποιήθηκε συλλογή δεδομένων. Συγκεκριμένα, ο κώδικας έτρεξε ασταμάτητα από την Παρασκευή 15 Σεπτεμβρίου 2023 (σύμφωνα με την κωδικοποίηση της συγκεκριμένης εργασίας 2023-09-16) στις 12:00 το πρωί έως την Πέμπτη 21 Σεπτεμβρίου 2023 (2023-09-21) στις 23:59. Στη συνέχεια, παρουσιάζονται γραφήματα και πίνακες προκειμένου να διευκολυνθεί η κατανόηση των αποτελεσμάτων, ενώ πραγματοποιείται επίσης έλεγχος για την εγκυρότητα του κώδικα.

9.2 Πλήθος γεγονότων ανά κατηγορία

Το πρώτο και κύριο γράφημα που αξίζει να αναφερθεί είναι το πλήθος των ειδήσεων ανά κατηγορία που καταγράφηκαν από το πληροφοριακό σύστημα, την χρονική περίοδο που αναφέρθηκε παραπάνω. Παρακάτω, παρουσιάζονται δύο γραφήματα για τη σύγκριση του πλήθους των γεγονότων ανα κατηγορία. Το πρώτο γράφημα, που προέρχεται από την πτυχιακή εργασία [Σ23], απεικονίζει το αρχικό πλήθος των γεγονότων ανα κατηγορία πριν την αφαίρεση των διπλοτύπων ειδησεογραφικών δεδομένων.

Σε παράλληλο πλαίσιο, δημιουργήθηκε το δεύτερο γράφημα, το οποίο απεικονίζει τη μείωση του αριθμού των γεγονότων κατά την ίδια περίοδο αφότου ανιχνεύθηκαν και αφαιρέθηκαν τα διπλότυπα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



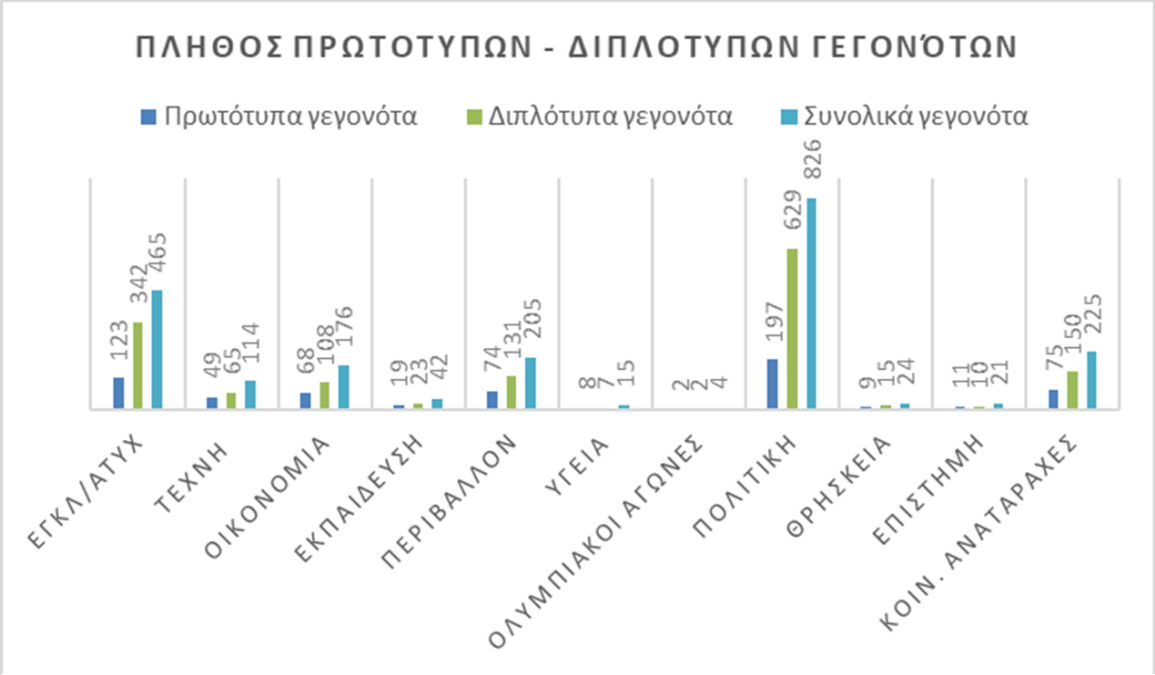
Γράφημα 9-1: Πλήθος γεγονότων ανά κατηγορία πριν και μετά την ανίχνευση διπλοτύπων

Παρατηρούμε πως ακόμα και μετά την αφαίρεση των διπλοτύπων, η φθίνουσα σειρά πλήθους γεγονότων παραμένει η ίδια. Δηλαδή, πρώτη εμφανίζεται η κατηγορία

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

politics με τα περισσότερα γεγονότα, που οφείλεται κυρίως στην διεξαγωγή δημοτικών και περιφερειακών εκλογών τον μήνα Οκτώβριο αλλά και στην εκλογή νέου προέδρου του κόμματος του ΣΥΡΙΖΑ. Έπειτα, εμφανίζονται με φθίνουσα σειρά οι κατηγορίες crime_accident, social_unrest, environment, economy, arts, education, science, religion, health και τέλος olympic_games.

Το παρακάτω γράφημα επικεντρώνεται στην εμφάνιση του πλήθους πρωτότυπων και διπλότυπων γεγονότων ανά κατηγορία σε σχέση με τα αρχικά δεδομένα.



Γράφημα 9-2: Πλήθος πρωτότυπων και διπλότυπων ανά κατηγορία

Για την επεξήγηση που παραπάνω γραφήματος παρουσιάζεται ο παρακάτω πίνακας που περιλαμβάνει και το ποσοστό του πλήθους των διπλότυπων γεγονότων.

Κατηγορία	Πρωτότυπες	Διπλότυπες	Συνολικές	Ποσοστό διπλοτύπων
Crime_accident	123	342	465	73,5%
Arts	49	65	114	57,01%
Economy	68	108	176	61,36%
Education	19	23	42	54,76%
Environment	74	131	205	63,90%
Health	8	7	15	46,67%
Olympic_games	2	2	4	50%

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Politics	197	629	826	79,15%
Religion	9	15	24	62,5%
Science	11	10	21	47,61%
Social_unrest	75	150	225	66,67%

Πίνακας 9-1: Πίνακας ποσοστού πλήθους διπλότυπων γεγονότων

Συγκεντρωτικά, λαμβάνοντας υπόψη και το γράφημα αλλά και τον πίνακα, γίνεται αντιληπτό πως η κατηγορία με τα περισσότερα διπλότυπα είναι η politics με ποσοστό 79,15% πιθανότατα λόγω των εκλογών για τον νέο πρόεδρο του ΣΥΡΙΖΑ αλλά και των εκλογών που θα πραγματοποιηθούν τον μήνα Οκτώβριο όσον αφορά τους δήμους και τις περιφέρειες. Ακολουθεί η κατηγορία crime_accident με ποσοστό 73,5% λόγω των δολοφονιών, τροχαίων και απωλειών λόγω της κακοκαιρίας στη Λιβύη που συνέβησαν την περίοδο αυτή. Στη συνέχεια βρίσκεται η κατηγορία social_unrest με ποσοστό διπλοτύπων 66,67% λόγω των απεργιών των ΜΜΜ αλλά και των πορειών που πραγματοποιήθηκαν ως προς τιμή του Παύλου Φύσσα. Ακολουθούν με φθίνουσα σειρά οι κατηγορίες environment, religion, economy, arts, education, olympic_games, science, health.

9.3 Γράφημα Word Cloud διπλοτύπων γεγονότων ανά κατηγορία

Ένα γράφημα Word Cloud, γνωστό και ως «tag cloud» ή «text cloud», είναι μια οπτική αναπαράσταση μιας συλλογής λέξεων, όπου το μέγεθος κάθε λέξης υποδεικνύει τη συχνότητα ή τη σημασία της εντός του κειμένου. Τα σύννεφα λέξεων χρησιμοποιούνται συνήθως για να απεικονίσουν τις πιο συχνές λέξεις σε ένα κείμενο ή σύνολο δεδομένων, κάνοντας τις λέξεις που εμφανίζονται πιο συχνά μεγαλύτερες και πιο εμφανείς [V23].

Σε αυτή την ανάλυση, χρησιμοποιούμε τα δεδομένα που απορρίφθηκαν ως διπλότυπα για να αποκτήσουμε μια εικόνα των γεγονότων που προσδιορίστηκαν ως σημαντικά από την κάθε κατηγορία γεγονότων. Με αυτόν τον τρόπο, μπορούμε να αποδείξουμε τους ισχυρισμούς που έγιναν στο ακριβώς παραπάνω υποκεφάλαιο, σχετικά με τον λόγο που εμφανίστηκε το μεγαλύτερο πλήθος διπλοτύπων

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

ειδησεογραφικών γεγονότων στις τρεις μεγαλύτερες – σε ποσοστό διπλοτύπων - κατηγορίες.

A)

B)

Γ)

Γράφημα 9-3: Word Cloud διπλοτύπων γεγονότων κατηγοριών Α) Εγκλήματα-Ατυχήματα, Β)Πολιτική, Γ)Κοινωνικές αναταραχές

Πράγματι, στην κατηγορία:

Α) *Εγκλήματα-Ατυχήματα (κωδικοποίηση κύριας κατηγορίας: 1),* προκύπτει ότι οι κυριότερες λέξεις που εμφανίζονται είναι οι «Λιβύη», «δολοφονία», και «τροχαίο». Η επικέντρωση της κατηγορίας εστιάζει κυρίως στην κατάσταση της Λιβύης μετά από την κακοκαιρία Daniel, όπου προκάλεσε μεγάλο αριθμό θυμάτων, αλλά και στο φονικό τροχαίο των Ελληνικών Ένοπλων Δυνάμεων, οι οποίοι είχαν αποσταλεί για βοήθεια στη Λιβύη, με αποτέλεσμα να υπάρχουν νεκροί και τραυματίες της Ελληνικής Αποστολής. Τέλος, αναφέρεται και η δολοφονία του Παύλου Φύσσα, που συζητείται ξανά μετά από χρόνια λόγω της επετείου του θανάτου του.

Β) *Πολιτική (κωδικοποίηση κύριας κατηγορίας: 11),* παρατηρούμε τις λέξεις «εκλογές», «ΣΥΡΙΖΑ», «Κασσελάκης», και «Αχτσιόγλου» να ξεχωρίζουν. Εδώ, η προσοχή επικεντρώνεται κυρίως στις εκλογές για τον νέο πρόεδρο του ΣΥΡΙΖΑ, μεταξύ των υποψηφίων Κασσελάκη Στέφανου και Αχτσιόγλου Έφης. Επίσης, συζητείται το θέμα των εκλογών που πρόκειται να διεξαχθούν τον επόμενο μήνα Οκτώβριο για τους δήμους και τις περιφέρειες.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Γ) *Κοινωνικές αναταραχές (κωδικοποίηση κύριας κατηγορίας: 4),* τονίζει τις λέξεις «απεργία», «ΟΗΕ», «πορεία», και «μετανάστες». Στο συνθετικό αυτό εγχείρημα, η προσοχή εστιάζεται κυρίως στην απεργία που οργανώθηκε από τα Μέσα Μαζικής Μεταφοράς (ΜΜΜ), στην Γενική Συνέλευση του ΟΗΕ, σε πορείες που πραγματοποιήθηκαν προς μνήμη του Παύλου Φύσσα, αλλά και στα ζητήματα που αφορούν τη μετανάστευση.

Αυτή η ανάλυση με τη χρήση των Word Cloud βοηθά να αποκτήσουμε μια βαθύτερη κατανόηση των σημαντικών γεγονότων σε κάθε κατηγορία και να επιβεβαιώσουμε τις παρατηρήσεις μας. Έτσι, μπορούμε να καταλάβουμε ποια γεγονότα ήταν τόσο σημαντικά ώστε να έχουν τόσα διπλότυπα και να δώσουμε ακόμα πιο εμπεριστατωμένες απαντήσεις στα ερωτήματα μας.

9.4 Γράφημα Word Cloud πρωτότυπων ειδήσεων ανά κατηγορία

Στη συνέχεια, εμφανίζονται τα word cloud που δημιουργήθηκαν από τις πρωτότυπες ειδήσεις για κάθε κατηγορία ξεχωριστά, εμφανίζοντας έτσι το μεγαλύτερο πλήθος της θεματολογίας για κάθε κατηγορία.

Γράφημα 9-4: Word cloud για τις κατηγορίες Α) Εγκλήματα-Ατυχήματα, Β)Περιβάλλον, Γ)Υγεία, Δ)Κοινωνικές αναταραχές

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Γράφημα 9-5: Word cloud για τις κατηγορίες Ε)Επιστήμη, Ζ)Τέχνη και Πολιτισμός, Η)Ολυμπιακοί αγώνες, Θ)Εκπαίδευση

Γράφημα 9-6: Word cloud για τις κατηγορίες Ι)Θρησκεία, Κ)Οικονομία, Λ)Πολιτική

Σύμφωνα με τα παραπάνω διαγράμματα word cloud θα αναλυθούν σε κάθε κατηγορία οι λέξεις που εμφανίζονται με την μεγαλύτερη συχνότητα. Συγκεκριμένα:

Α) *Εγκλήματα-Ατυχήματα (κωδικοποίηση κύριας κατηγορίας: 1):* «Λιβύη», «δολοφονία», «μέτρα», «Μητσοτάκης». Φαίνεται πως αυτή η κατηγορία επικεντρώνεται κυρίως στην κατάσταση της Λιβύης μετά από την κακοκαιρία Daniel όπου στο πέρασμά της άφησε μεγάλο πλήθος θυμάτων, στην δολοφονία του Παύλου

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Φύσσα που συζητείται ξανά μετά από χρόνια λόγω της επετείου του θανάτου του, όπως και του 36χρονου Αντώνη Καρυώτη στο «Blue Horizon» από το πλήρωμα του επονομαζόμενου πλοίου. Τέλος, εστιάζει στην ανακοίνωση ορισμένων μέτρων από τον πρωθυπουργό της χώρας, Κυριάκο Μητσοτάκη.

Β) *Περιβάλλον (κωδικοποίηση κύριας κατηγορίας: 2):* «Θεσσαλία», «πλημμύρες», «Λιβύη», «φωτιά». Εστιάζει κυρίως στις πλημμύρες που έπληξαν την Θεσσαλία και την Λιβύη, όπου προκάλεσαν πλήθος καταστροφών αλλά και σε περιστατικά φωτιάς που είχαν προκύψει παράλληλα αυτή την περίοδο σε διάφορες περιοχές της Ελλάδας.

Γ) *Υγεία (κωδικοποίηση κύριας κατηγορίας: 3):* «εμβόλιο», «moderna», «επιστήμονες». Αυτή η κατηγορία φαίνεται να επικεντρώνεται κυρίως στην επιστημονική ανακάλυψη ενός αντιγριπικού εμβολίου της εταιρείας Moderna.

Δ) *Κοινωνικές αναταραχές (κωδικοποίηση κύριας κατηγορίας: 4):* «απεργία», «ΟΗΕ». Επικεντρώνεται κυρίως στην απεργία που έκαναν στα ΜΜΜ και στην Γενική Συνέλευση του ΟΗΕ όπου παραστάθηκαν πολλοί ηγέτες χωρών.

Ε) *Επιστήμη (κωδικοποίηση κύριας κατηγορίας: 5):* «moderna», «αντιγριπικό», «εμβόλιο», «ανοσολογική». Επικεντρώνεται στην ανακάλυψη ενός αντιγριπικού εμβολίου από την εταιρία Moderna που αποσκοπεί στην καλύτερη ανοσολογική προστασία.

Ζ) *Τέχνη και Πολιτισμός (κωδικοποίηση κύριας κατηγορίας: 6):* «ΔΕΘ», «Μητσοτάκης», «συνέντευξη». Εστιάζει κυρίως στην ΔΕΘ (Διεθνής Έκθεση Θεσσαλονίκης), στις συνεντεύξεις που έδωσε ο Μητσοτάκης εκεί αλλά και σε συνεντεύξεις άλλων σημαντικών προσώπων.

Η) *Ολυμπιακοί αγώνες (κωδικοποίηση κύριας κατηγορίας: 7):* «διεθνής», «διάκριση». Επικεντρώνεται κυρίως στην διάκριση της Ελληνικής ομάδας κολύμβησης στους 3ους Μεσογειακούς Παράκτιους Αγώνες.

Θ) *Εκπαίδευση (κωδικοποίηση κύριας κατηγορίας: 8):* «Στέφανος Κασσελάκης», «Θεσσαλία», «σχολεία». Αυτή η κατηγορία εστιάζει κυρίως στην εκπαιδευτική διαδικασία που πρόκειται να υιοθετηθεί στη Θεσσαλία μετά τις καταστροφικές πλημμύρες που προκάλεσε η κακοκαιρία Daniel, καθώς και δηλώσεις του Κασσελάκη σχετικά με θέματα εκπαίδευσης πριν τις εκλογές για τη θέση του προέδρου του κόμματος Σύριζα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Ι) *Θρησκεία (κωδικοποίηση κύριας κατηγορίας: 9):* «Λιβύη», «Άγιον όρος», «μοναχός», «συνελήφθη». Αυτή η κατηγορία επικεντρώνεται κυρίως στις θρησκευτικές διαδικασίες, ειδικότερα στις τελετές κηδειών, που ακολουθεί η Λιβύη λόγω της τραγωδίας που προκλήθηκε από την κακοκαιρία Daniel. Επιπλέον, αναφέρεται σε ένα συγκεκριμένο περιστατικό στο Άγιον Όρος, όπου ένας μοναχός συνελήφθη λόγω της εμπλοκής του σε ένα περιστατικό βίας εναντίον ενός άλλου μοναχού.

Κ) *Οικονομία (κωδικοποίηση κύριας κατηγορίας: 10):* «Θεσσαλία», «τράπεζες», «επιχειρήσεις». Εστιάζει κυρίως στις καταστροφές των επιχειρήσεων στη Θεσσαλία από την κακοκαιρία Daniel. Επιπλέον αναφέρονται και γεγονότα σχετικά με τις αλλαγές που γίνονται σε τραπεζικά ζητήματα, όπως επιτόκια και μεταβολές του χρηματιστηρίου.

Λ) *Πολιτική (κωδικοποίηση κύριας κατηγορίας: 11):* «Μητσοτάκης», «Θεσσαλία», «ΔΕΘ», «Κασσελάκης». Επικεντρώνεται κυρίως στις δηλώσεις που έκανε ο Μητσοτάκης στην ΔΕΘ για την καταστροφή στη Θεσσαλία από τις πλημμύρες αλλά και στις δηλώσεις που έκανε ο Κασσελάκης σχετικά με τις εκλογές για τον νέο πρόεδρο του κόμματος Σύριζα.

9.5 Πλήθος διπλότυπων και πρωτότυπων ειδήσεων ανά ημέρα

Στο παρακάτω γράφημα εμφανίζεται η διαφορά μεταξύ των διπλότυπων σε σχέση με τις πρωτότυπες ειδήσεις ανά ημέρα καταγραφής των γεγονότων.

Γράφημα 9-7: Πλήθος διπλότυπων και πρωτότυπων ειδήσεων ανά ημέρα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

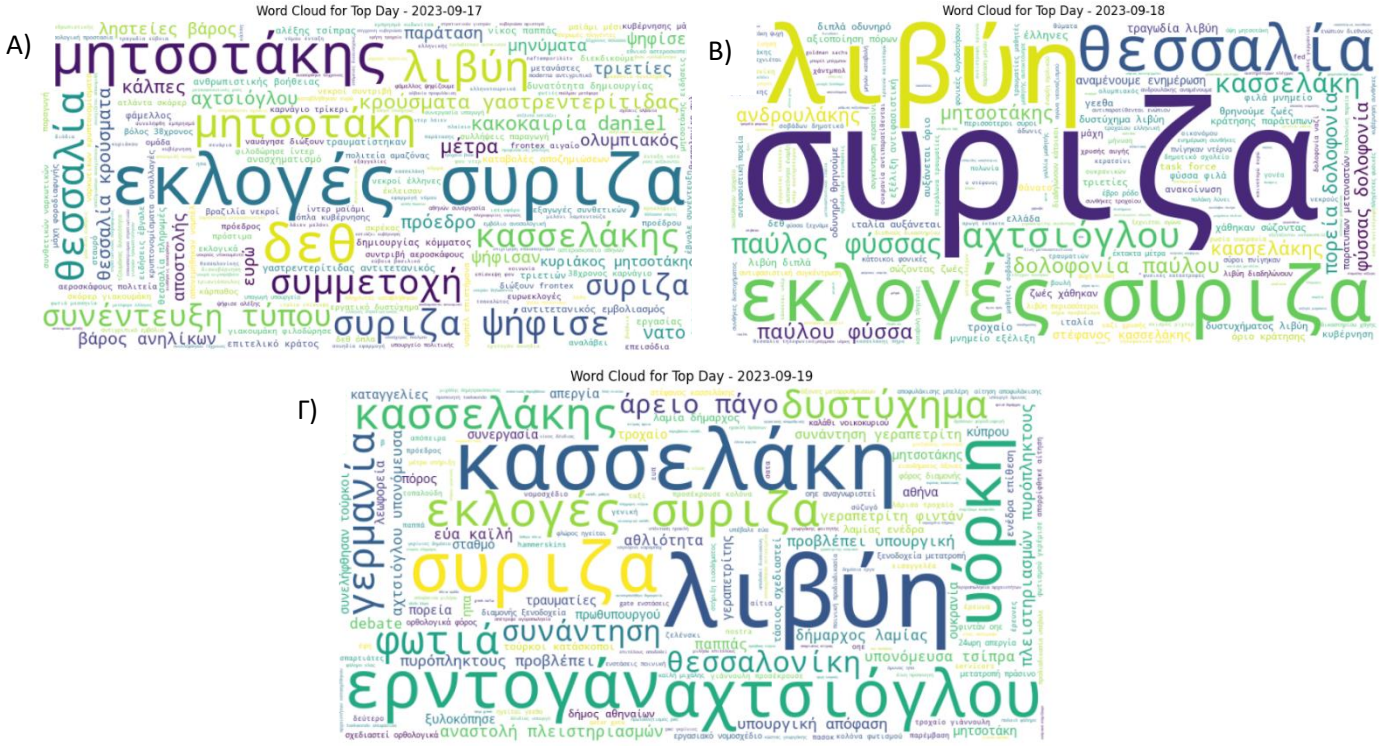
Η μεγαλύτερη διαφορά μεταξύ πρωτότυπων και διπλότυπων ειδήσεων γίνεται την Κυριακή 2023-09-17 όπου υπάρχει η διαφορά των 133 ειδήσεων. Έπειτα γίνεται την Δευτέρα 2023-09-18 όπου υπάρχει η διαφορά των 111 ειδήσεων. Η ημέρα Παρασκευή 2023-09-15 δεν μπορεί να συγκριθεί δικαίως με τις υπόλοιπες διότι ξεκίνησε να γίνεται συλλογή δεδομένων την Παρασκευή από τις 12 το μεσημέρι και έπειτα.

Η διαφορά μεταξύ των πρωτότυπων και διπλότυπων ειδήσεων ανά ημέρα καταγραφής των γεγονότων μπορεί να δείχνει πολλά πράγματα. Ορισμένες πιθανές ερμηνείες των παραπάνω αποτελεσμάτων είναι:

1. Δημοφιλία των ειδησεογραφικών γεγονότων: Η διαφορά στον αριθμό των ειδήσεων μεταξύ των πρωτότυπων και των διπλότυπων ειδήσεων ανά ημέρα μπορεί να αντανακλά τη δημοφιλία των ειδήσεων για συγκεκριμένες ημερομηνίες. Κατά τη διάρκεια του Σαββατοκύριακου (17 και 18 Σεπτεμβρίου 2023) και της Δευτέρας (19 Σεπτεμβρίου 2023), πιθανόν να συνέβησαν γεγονότα που προκάλεσαν μεγάλο ενδιαφέρον, επομένως, οι ειδήσεις σε αυτές τις ημερομηνίες ήταν πιο πολλές και επικυρωμένες.

   Πράγματι, αν εξετάσουμε τα Word Cloud που δημιουργούνται από τα δεδομένα -πρωτότυπα και διπλότυπα γεγονότα μαζί - τις ημέρες εκείνες, μπορούμε να παρατηρήσουμε πως τα κύρια γεγονότα που υπέστησαν μεγάλη κάλυψη από τις ειδησεογραφικές ιστοσελίδες ήταν οι εκλογές για τον νέο πρόεδρο του ΣΥΡΙΖΑ, οι καταστροφές στη Θεσσαλία από τη κακοκαιρία Daniel και τέλος η Λιβύη παράλληλα για τις καταστροφές που υπέστη από την κακοκαιρία Daniel αλλά και για το δυστύχημα των ελληνικών ένοπλων δυνάμεων στη χώρα αυτή. Αυτά τα γεγονότα φαίνεται να είχαν τη μεγαλύτερη κάλυψη και προσοχή από τα μέσα ενημέρωσης κατά τη διάρκεια των συγκεκριμένων ημερών.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38



Γράφημα 9-8: Word cloud των ημερομηνιών: Α) 17-09-2023, Β) 18-09-2023, Γ) 19-09-2023

2. Επίδραση των ημερών της εβδομάδας: Οι διαφορές μεταξύ των ημερών μπορεί να οφείλονται στην επίδραση των ημερών της εβδομάδας στη δημοφιλία των ειδήσεων. Συνήθως, τα Σαββατοκύριακα είναι ημέρες όπου πολλοί άνθρωποι έχουν περισσότερο χρόνο να διαβάσουν ειδήσεις και ενδέχεται να αντιδρούν περισσότερο σε γεγονότα, οπότε ίσως επιλέγεται συνειδητά από τους δημοσιογράφους η δημοσίευση περισσότερων ειδησεογραφικών γεγονότων τις ημέρες εκείνες.
3. Δημοφιλία της Δευτέρας: Συχνά, η Δευτέρα θεωρείται μια ημέρα που σηματοδοτεί την αρχή μιας νέας εβδομάδας. Το ενδιαφέρον για ειδήσεις μπορεί να αυξηθεί μετά το Σαββατοκύριακο, όταν πολλοί άνθρωποι επιστρέφουν στην καθημερινή ρουτίνα τους. Επομένως, τη Δευτέρα είναι πιθανό να υπάρξει αυξημένο ενδιαφέρον για τις ειδήσεις και να καταγράφονται περισσότερες πρωτότυπες ειδήσεις.

Συνολικά, η διαφορά αυτή μπορεί να οδηγήσει σε συμπεράσματα σχετικά με τη δημοφιλία των ειδήσεων σε συγκεκριμένες ημερομηνίες και την επίδραση των ημερών της εβδομάδας στη δημοσιογραφία και την αναγνώριση των ειδησεογραφικών γεγονότων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

9.6 Μετρικές αξιοπιστίας του συστήματος για την εύρεση διπλοτύπων ειδησεογραφικών γεγονότων

Κρίσιμη κρίνεται η αξιοπιστία τους πληροφοριακού συστήματος όσον αφορά την αναγνώριση των διπλοτύπων. Για τον λόγο αυτό έγινε χειροκίνητα η αναγνώριση των διπλοτύπων σύμφωνα με τον τρόπο υλοποίησης του κώδικα. Συγκεκριμένα, χρησιμοποιήθηκε η κατηγορία crime_accident που έχει πλήθος ειδήσεων 465 συμπεριλαμβανομένων και γεγονότων των οποίων κατηγοριοποιήθηκαν από τον πίνακα των duplicates. Ο έλεγχος όλων των γεγονότων από όλες τις κατηγορίες ήταν αρκετά χρονοβόρος λόγω του μεγάλου πλήθους των γεγονότων. Έτσι, η επιλογή της συγκεκριμένης κατηγορίας θεωρήθηκε ως ένα αντιπροσωπευτικό δείγμα, το οποίο σύμφωνα με το πληροφοριακό σύστημα περιείχε ένα μεγάλο αριθμό διπλοτύπων ειδήσεων, όπως επιβεβαιώθηκε και από το σχήμα του γραφήματος 9-7.

Πριν αναφερθούμε στην διαδικασία που ακολουθήθηκε, αξίζει να σημειωθεί πως ο σκοπός της παρούσας εργασίας είναι η εξέταση του συνόλου δεδομένων που προέκυψε από την εργασία της συναδέλφου μου Σ. Σκορδίλη [Σ23], με σκοπό τον εντοπισμό των διπλοτύπων ειδήσεων. Παρατηρήθηκε ότι το αρχικό σύνολο δεδομένων που χρησιμοποιήθηκε για την ανάλυση των ειδήσεων, και το οποίο είχε προκύψει με πλήρως αυτοματοποιημένη τεχνική από την εργασία [Σ23] περιείχε κάποια λίγα σφάλματα στην αναγνώριση των γεγονότων. Αυτά τα λίγα σφάλματα μπορεί να περιλαμβάνουν και false positives, δηλαδή ειδήσεις που αναγνωρίστηκαν ως γεγονότα όμως δεν ήταν στην πραγματικότητα γεγονότα, με βάση τον ορισμό που είχε καθοριστεί.

Παρά το γεγονός ότι το αρχικό σύνολο δεδομένων περιείχε αυτά τα λίγα σφάλματα στην αναγνώριση των γεγονότων, η παρούσα εργασία επικεντρώνεται στον ακριβή εντοπισμό των επαναλαμβανόμενων ειδήσεων, ανεξάρτητα από την οποιαδήποτε τυχόν ανακρίβεια. Η πρωταρχική διερεύνηση αφορά τον εντοπισμό αυτών των διπλοτύπων ειδήσεων και δεν στοχεύει στην διόρθωση σφαλμάτων που μπορεί να περιέχει το αρχικό σύνολο δεδομένων. Παρόλα αυτά, ακόμη και ένα μικρό σφάλμα μπορεί να επηρεάσει την ακρίβεια των αποτελεσμάτων της παρούσας εργασίας.

Επομένως, χρησιμοποιώντας το σύνολο δεδομένων που παρήχθη από την εργασία [Σ23], στη συνέχεια θα προβούμε στην εύρεση των πρωτότυπων γεγονότων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Για την απόφαση αυτή έγινε χρήση της συνάρτησης ΕύρεσηΠρωτοτύπων(Χ), όπου Χ είναι ένα γεγονός. Επομένως, για την αποδοτικότητα της συνάρτησης αυτής θα καταμετρηθούν τα TP, TN, FP και FN. Οπότε, ο έλεγχος των πρωτότυπων ειδήσεων γίνεται, με βάση τα παρακάτω κριτήρια:

1. Η βάση δεδομένων στην αρχή είναι κενή, άρα για τον άνθρωπο-παρατηρητή το γεγονός Α είναι πρωτότυπο, όχι διπλότυπο. Κατέταξε η συνάρτηση ΕύρεσηΠρωτοτύπων(Α) το Α στα πρωτότυπα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε TP ενώ αν η απάντηση είναι αρνητική, τότε έχουμε FN.
2. Για τον άνθρωπο-παρατηρητή το γεγονός Β είναι πρωτότυπο, καθώς πραγματεύεται διαφορετικό γεγονός σε σχέση με το Α. Κατέταξε η συνάρτηση το γεγονός Β στα πρωτότυπα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε TP ενώ αν η απάντηση είναι αρνητική, τότε έχουμε FN.
3. Για τον ίδιο άνθρωπο-παρατηρητή το γεγονός Γ είναι διπλοεγγραφή. Κατέταξε η συνάρτηση ΕύρεσηΠρωτοτύπων(Γ) το γεγονός Γ στα πρωτότυπα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε FP ενώ αν η απάντηση είναι αρνητική, τότε έχουμε TN. Για την πραγματοποίηση των παραπάνω, η διαδικασία που χρησιμοποιήθηκε ήταν η ταξινόμηση όλων των ειδήσεων βάσει id, διότι η είδηση με το μικρότερο id ήταν αυτή που εισήχθη στη βάση δεδομένων πρώτη.

Ακολουθώντας επιμελώς τα παραπάνω, τα αποτελέσματα που προκύπτουν, φαίνονται στην συνέχεια:

- TP = 105
- TN = 277
- FP = 18
- FN = 65

Εάν τα αρχικά δεδομένα περιέχουν ανακρίβειες ή είναι ελλιπή, μπορούν να επηρεάσουν την ακρίβεια και τις μετρικές αξιολόγησης, ιδίως τον αριθμό των FN. Οι ψευδείς αρνητικοί είναι τα πραγματικά πρωτότυπα γεγονότα που το σύστημα αναγνωρίζει εσφαλμένα ως διπλότυπα. Έχοντας αυτά τα στοιχεία, είναι εφικτός ο υπολογισμός μέτρων αξιολόγησης. Ειδικότερα:

- **Ορθότητα (Accuracy)** = (TP+TN)/(TP+TN+FP+FN) = 0,82

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Η ακρίβεια αναφέρεται στο ποσοστό των ορθών προβλέψεων σε σχέση με το σύνολο των προβλέψεων. Σε αυτήν την περίπτωση, το Accuracy είναι ίσο με 0,82, πράγμα που υποδηλώνει ότι ένα ποσοστό 82% των προβλέψεων είναι σωστές.

- **Ανάκληση (Recall)** = TP/(TP+FN) = 0,62

  Η ανάκληση αναφέρεται στο ποσοστό των πραγματικών θετικών περιπτώσεων που αναγνωρίστηκαν από το σύστημα. Το Recall είναι 0,62, που σημαίνει ότι το σύστημα αναγνώρισε το 62% των πραγματικών θετικών περιπτώσεων.

- **Ακρίβεια (Precision)** = TP/(TP+FP) = 0,85

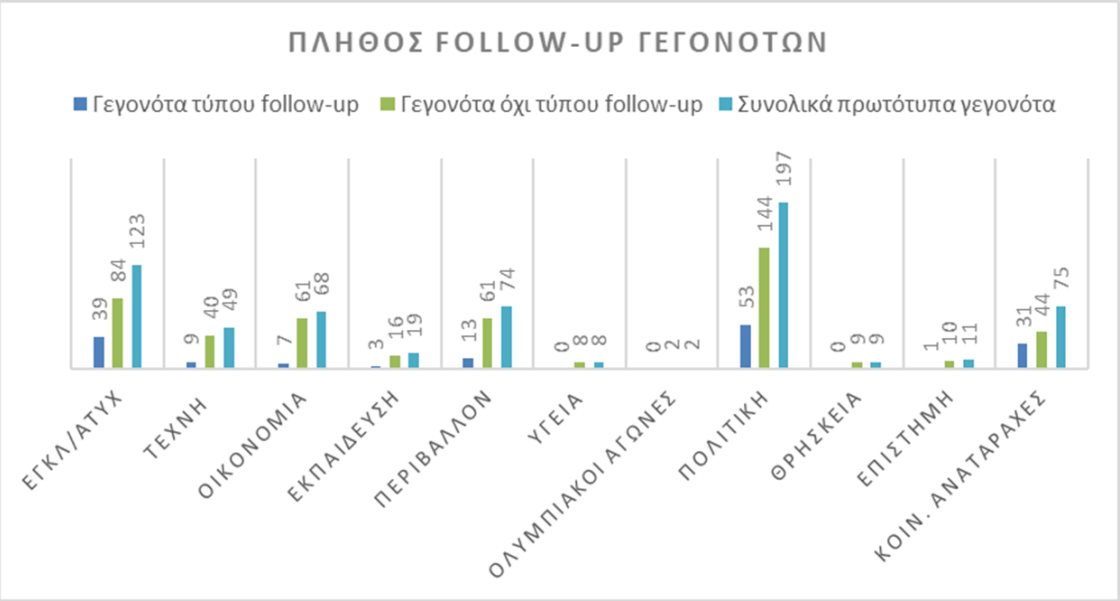
  Η ακρίβεια αναφέρεται στο ποσοστό των προβλέψεων που είναι πραγματικά θετικές. Η μετρική Precision είναι 0,85, προσδιορίζοντας ότι το σύστημα είναι αρκετά καλό στο να μην αναγνωρίζει αρνητικά περισσότερες περιπτώσεις από αυτές που είναι πραγματικά θετικές.

- **F1-score** = 2*TP/ (2*TP + FP + FN) = 0,72

  Το F1-Score είναι ένας συνδυασμός της Precision και της Recall και μετρά την ακρίβεια των θετικών προβλέψεων συνολικά. Έχει τιμή 0,72, υποδηλώνοντας ένα ικανοποιητικό συνολικό επίπεδο ακρίβειας.

9.7 Πλήθος follow up ανά κατηγορία

Το παρακάτω γράφημα παρουσιάζει το πλήθος των follow up γεγονότων που ανιχνεύθηκαν ανά κατηγορία.



Γράφημα 9-9: Πλήθος follow up ανά κατηγορία

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Παρατηρείται πως η κατηγορία politics έχει τις περισσότερες follow up ειδήσεις σε ποσοστό 26,9% από τις 197 ειδήσεις που αποτελείται. Αυτό οφείλεται στην διεξαγωγή δημοτικών εκλογών τον μήνα Οκτώβριο αλλά και στην εκλογή νέου προέδρου του κόμματος του ΣΥΡΙΖΑ. Συγκεκριμένα, μετά από ανασκόπηση των ειδήσεων, έγινε αντιληπτό πως οι περισσότερες αναφέρονται στις επώνυμες οντότητες «Κασσελάκης», «Σύριζα», «Περιφέρεια» και «Αχτσιόγλου» ενημερώνοντας συνεχώς για τις εξελίξεις των εκλογών.

Η αμέσως επόμενη κατηγορία είναι το crime_accident σε ποσοστό 31,7% από τις 123 ειδήσεις που αποτελείται. Μετά από επανεξέταση των ειδήσεων, η πλειονότητα των επώνυμων οντοτήτων αναφέρεται στο «Blue Horizon» που αντιστοιχεί στην είδηση της υπόθεσης δολοφονίας του 36χρονου Αντώνη Καρυώτη από το πλήρωμα του επονομαζόμενου πλοίου. Οι follow up ειδήσεις σχετίζονται με την απαγγελία κατηγοριών και την εξέλιξη της υπόθεσης. Ακόμη, πολύ συχνά εμφανίζεται η «Λιβύη» όπου η Ελλάδα έστειλε βοήθεια για τις καταστροφικές πλημμύρες που έλαβαν χώρα εκείνο το διάστημα όμως, μετά από τραγικό δυστύχημα, έχασαν τη ζωή τους πέντε και τραυματίστηκαν δεκατρία μέλη της ελληνικής αποστολής βοήθειας. Επομένως, οι follow up ειδήσεις αναφέρονται στην τροπή των καταστροφών, στην έρευνα για τις συνθήκες του δυστυχήματος και στην κατάσταση των επιζώντων.

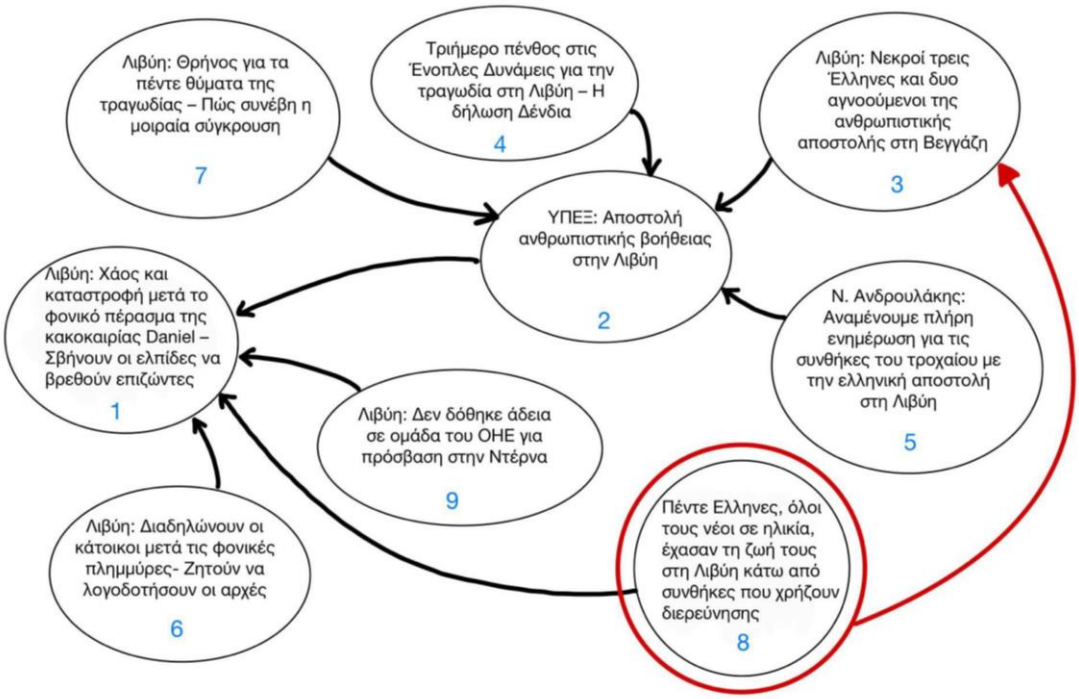
Η τελευταία κατηγορία με μεγάλο πλήθος follow up είναι η social_unrest σε ποσοστό 41,3% από τις 75 ειδήσεις που αποτελείται. Αυτό οφείλεται κυρίως στην επέτειο θανάτου του Παύλου Φύσσα και των πορειών που γίνονται κάθε χρονιά εις μνήμης του, διότι εμφανίζεται πολύ συχνά η οντότητα «Φύσσας». Επιπλέον, παρατηρείται μεγάλο πλήθος ειδήσεων για την απεργία των μέσων μαζικής μεταφοράς με τις επώνυμες οντότητες «ΜΜΜ» και «Μετρό». Τέλος, εμφανίζεται η οντότητα «Λιβύη» λόγω των πολυάριθμων διαδηλώσεων από τους πολίτες στις πληγείσες περιοχές από τις πλημμύρες.

Μπορούμε να παρατηρήσουμε πως τα γεγονότα που παρουσιάζουν μεγάλο πλήθος follow-up ειδήσεων ταυτίζονται κατά κύριο λόγο με τα γεγονότα που παρουσίασαν μεγάλο αριθμό διπλοτύπων. Ένα γεγονός που παράγει πολλά διπλότυπα και έχει πολλές εξελίξεις είναι πιθανόν να θεωρείται σημαντικό. Αυτό μπορεί να σημαίνει ότι το γεγονός έχει ευρεία κάλυψη στα μέσα ενημέρωσης ή ότι έχει σημαντικές επιπτώσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

9.8 Σχεσιακός Γράφος γεγονότος follow-up είδησης

Στην υποενότητα αυτή, παρουσιάζεται η εξέλιξη ενός γεγονότος σε έναν γράφο σύμφωνα με τον τρόπο που το κατέγραψε το συγκεκριμένο πληροφοριακό σύστημα.



Γράφος 9-1: Σχεσιακός Γράφος follow-up ειδήσεων για τα γεγονότα στη Λιβύη

Στον παραπάνω γράφο απεικονίζεται η εξέλιξη των γεγονότων στην Λιβύη, σύμφωνα με το προτεινόμενο σύστημα. Η Λιβύη, κατά το χρονικό διάστημα που συλλέχθηκαν τα δεδομένα για την πτυχιακή εργασία, βρέθηκε αντιμέτωπη με σημαντικές καταστροφές λόγω των επικίνδυνων πλημμύρων που προκάλεσε η κακοκαιρία Daniel. Επιπλέον, οι Ελληνικές Ένοπλες Δυνάμεις ανακοίνωσαν την αποστολή βοήθειας στη Λιβύη για την αντιμετώπιση των καταστροφών, όμως δυστυχώς, συνέβη ένα τροχαίο ατύχημα με αποτέλεσμα να υπάρχουν νεκροί και τραυματίες μεταξύ των μελών της Ελληνικής Αποστολής. Παράλληλα με αυτά, οι κάτοικοι της Λιβύης εξέφρασαν τη δυσαρέσκειά τους με διαδηλώσεις, λόγω των καταστροφών που προκλήθηκαν από την κακοκαιρία.

Ο γράφος λοιπόν αναφέρει γεγονότα σχετικά με τις τρεις καταστάσεις που προαναφέρθηκαν βρίσκοντας την μητρική αλλά και γεγονότα που είναι εξέλιξη αυτής. Οι ειδήσεις αυτές καταγράφθηκαν στο χρονικό πλαίσιο 16-09-2023 έως 19-09-2023. Οι αριθμοί με μπλε χρώμα αναδεικνύουν τη σειρά με την οποία δημοσιεύτηκαν τα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

συγκεκριμένα ειδησεογραφικά γεγονότα. Με κόκκινο αναγράφεται μία είδηση, που αναφέρεται στο τροχαίο της Ελληνικής Αποστολής και έχει βρεθεί ως follow-up είδηση των πλημμυρών στη Λιβύη. Αν και όχι λάθος από την έννοια της εξέλιξης του γεγονότος αυτού, διότι χωρίς τις καταστροφές στη Λιβύη δε θα είχε σταλθεί βοήθεια από τις Ελληνικές Δυνάμεις στη Λιβύη, το σύστημα θα ήταν πιο ακριβές αν «έδειχνε» στην είδηση του τροχαίου.

Το σύστημα δηλαδή, δημιουργεί μια μικρή ανακριβή προτεραιότητα. Εάν βρεθεί μια αντιστοιχία μεταξύ των γεγονότων, συνδέει το πρόσφατο γεγονός με το πρώτο που βρέθηκε κατά την αναζήτηση. Το σύστημα εξετάζει τρεις ημέρες πίσω από το κάθε γεγονός. Αυτή η διαδικασία επιτρέπει στο σύστημα να εντοπίζει πιθανές εξελίξεις και συσχετίσεις μεταξύ γεγονότων. Παρόλο που αυτή η μέθοδος είναι χρήσιμη για τον εντοπισμό πιθανών συνδέσεων, είναι σημαντικό να σημειώσουμε ότι μπορεί να υπάρχουν ανακρίβειες στον τρόπο με τον οποίο συνδέονται τα γεγονότα. Ειδικότερα, το σύστημα μας εντοπίζει αντιστοιχίες με βάση την περίοδο προηγούμενων ημερών, και όχι απαραίτητα βάσει της ακριβούς χρονικής σειράς των γεγονότων.

Βέβαια, υπήρχαν και άλλα γεγονότα σχετικά με την Λιβύη όμως δεν αντιστοιχήθηκαν σε κάποια από τις δύο επικρατούσες καταστάσεις, γι'αυτό δεν απεικονίζονται στον γράφο.

Καταλήγοντας, αξίζει να σημειωθεί πως, το αρχικό σύνολο δεδομένων περιέχει κάποια λίγα σφάλματα όσον αφορά τα γεγονότα με τον αυστηρό ορισμό που δόθηκε στην πτυχιακή εργασία [Σ23]. Παρά το γεγονός ότι το αρχικό σύνολο δεδομένων ανέφερε σφάλματα στην αναγνώριση των γεγονότων, η εργασία επικεντρώνεται στον εντοπισμό των follow-up ειδήσεων, ανεξάρτητα από τυχόν ανακρίβειες. Έτσι, μπορεί στον γράφο να υπάρχουν κάποιες πολύ λίγες ειδήσεις που δεν ανταποκρίνονται στον ορισμό που δόθηκε για τα γεγονότα.

9.9 Μετρικές αξιοπιστίας του συστήματος για την εύρεση follow up ειδησεογραφικών γεγονότων

Για την επαλήθευση της σωστής λειτουργίας του πληροφοριακού συστήματος, πραγματοποιήθηκε χειρωνακτική διαδικασία, σύμφωνα με τις προδιαγραφές του κώδικα, για την εύρεση αν ένα γεγονός είναι πρωτότυπο ή εξαρτάται από κάποιο άλλο υποδηλώνοντας έτσι πως υπάρχει εξέλιξη του πρωτότυπου γεγονότος. Για τον έλεγχο

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

αυτό επιλέχθηκε μόνο η κατηγορία crime_accident που έχει πλήθος 123 ειδήσεων. Ο έλεγχος όλων των γεγονότων από όλες τις κατηγορίες ήταν αρκετά χρονοβόρος λόγω του μεγάλου πλήθους των γεγονότων. Παρόλα αυτά, έγινε η επιλογή της συγκεκριμένης κατηγορίας διότι είχε αρκετά γεγονότα που ήταν εξέλιξη της πρωτότυπής τους είδησης (το χρονικό διάστημα 15-21 Σεπτεμβρίου 2023, όπου χρησιμοποιήθηκαν οι ειδήσεις) άρα τα αποτελέσματα είναι πιο έγκυρα από ότι αν γινόταν επιλογή κάποιας άλλης κατηγορίας.

Πριν αναφερθούμε στη διαδικασία που ακολουθήθηκε, αξίζει να σημειώσουμε ότι ο στόχος της παρούσας εργασίας είναι η εξέταση του συνόλου δεδομένων που προέκυψε από την εργασία [Σ23], με σκοπό τον εντοπισμό των follow-up ειδήσεων. Παρατηρήθηκε ότι το αρχικό σύνολο δεδομένων, που χρησιμοποιήθηκε για την ανάλυση των ειδήσεων, περιείχε πιθανές ανακρίβειες στην αναγνώριση των γεγονότων. Αυτές οι ανακρίβειες μπορεί να περιλαμβάνουν και false positives, δηλαδή ειδήσεις που αναγνωρίστηκαν ως γεγονότα, χωρίς ωστόσο να ανταποκρίνονται στον ορισμό που καθορίστηκε.

Παρά τις ανακρίβειες που ενδέχεται να περιέχει το αρχικό σύνολο δεδομένων, η εργασία επικεντρώνεται στον ακριβή εντοπισμό των ειδήσεων που ήταν εξέλιξη κάποιας άλλης είδησης, ανεξάρτητα από τυχόν ανακρίβειες. Η πρωταρχική διερεύνηση αφορά τον εντοπισμό αυτών των follow-up ειδήσεων και δεν στοχεύει στη διόρθωση των ανακριβειών που μπορεί να περιέχει το αρχικό σύνολο δεδομένων. Παρόλα αυτά, αξίζει να σημειώσουμε ότι οι εν λόγω ανακρίβειες μπορεί να επηρεάσουν την ακρίβεια των αποτελεσμάτων της παρούσας εργασίας.

Επομένως, χρησιμοποιώντας το σύνολο δεδομένων της εργασίας [Σ23], θα προβούμε στην εύρεση των μη follow-up ειδήσεων. Για την απόφαση αυτή έγινε χρήση της συνάρτησης ΕύρεσηΜηΕξαρτημένωνΓεγονότων(Χ) όπου Χ, είναι ένα πρωτότυπο γεγονός. Κατά τα γνωστά, για την αποδοτικότητα της συνάρτησης αυτής θα καταμετρηθούν τα TP, TN, FP και FN. Επομένως, ο έλεγχος των μη follow up γεγονότων γίνεται σύμφωνα με τα παρακάτω κριτήρια:

1. Η βάση δεδομένων στην αρχή είναι κενή, άρα για τον άνθρωπο-παρατηρητή το γεγονός Α προφανώς δεν είναι follow up. Εντόπισε η συνάρτηση ΕύρεσηΜηΕξαρτημένωνΓεγονότων (Α) ότι το γεγονός Α δεν είναι εξαρτημένο

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

από άλλο προγενέστερο γεγονός; Αν η απάντηση είναι θετική, τότε έχουμε TP ενώ αν η απάντηση είναι αρνητική, τότε έχουμε FN.

2. Για τον άνθρωπο-παρατηρητή το γεγονός B αποτελεί follow up του γεγονότος A. Κατέταξε η συνάρτηση ΕύρεσηΜηΕξαρτημένωνΓεγονότων(B) το γεγονός B στα μη εξαρτημένα από άλλα προγενέστερα γεγονότα; Αν η απάντηση είναι θετική, τότε έχουμε FP ενώ αν η απάντηση είναι αρνητική, τότε έχουμε TN.

Εφαρμόζοντας τα παραπάνω κριτήρια με προσοχή, προέκυψαν τα ακόλουθα αποτελέσματα:

- TP = 81
- TN = 7
- FP = 12
- FN = 23

Με τα αποτελέσματα αυτά, υπάρχει δυνατότητα υπολογισμού των μετρικών. Αναλυτικότερα:

- **Ορθότητα (Accuracy)** = (TP+TN)/(TP+TN+FP+FN) = 0,715

  Αυτή η μετρική δείχνει πόσο καλά το σύστημα αναγνωρίζει τα γεγονότα ως εξέλιξη κάποιου άλλου γεγονότος ή όχι. Υπολογίζεται ως τον λόγο των σωστών προβλέψεων (TP και TN) προς τον συνολικό αριθμό των προβλέψεων. Η ορθότητα είναι 0,715, που αντιστοιχεί σε περίπου 72% ορθότητα.
- **Ανάκληση (Recall)** = TP/(TP+FN) = 0,778

  Αυτή η μετρική δείχνει πόσα από τα πραγματικά όχι follow-up γεγονότα αναγνωρίστηκαν σωστά από το σύστημα. Υπολογίζεται ως τον λόγο των σωστών όχι follow-up προβλέψεων (TP) προς τον συνολικό αριθμό των όχι follow-up γεγονότων. Η ανάκληση υπολογίστηκε ως 0,778, που σημαίνει ότι το σύστημα αναγνώρισε το ~78% των πραγματικών όχι follow-up περιπτώσεων.
- **Ακρίβεια (Precision)** = TP/(TP+FP) = 0,871

  Αυτή η μετρική δείχνει πόσα από τα γεγονότα που το σύστημα αναγνωρίζει ως εξαρτημένα πράγματι είναι εξαρτημένα. Υπολογίζεται ως τον λόγο των σωστών όχι follow-up προβλέψεων (TP) προς τον συνολικό αριθμό των οχι follow-up προβλέψεων (TP και FP). Η ακρίβεια είναι 0,871, που αντιστοιχεί σε περίπου 87% ακρίβεια, προσδιορίζοντας ότι το σύστημα είναι αρκετά καλό στο

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

να μην αναγνωρίζει αρνητικά περισσότερες περιπτώσεις από αυτές που είναι πραγματικά θετικές.

- **F1-score** = 2*TP/ (2*TP + FP + FN) = 0,822

  Το F1-Score είναι ένας συνδυασμός της ανάκλησης και της ακρίβειας και δίνει μια ιδέα για τη συνολική επίδοση του συστήματος. Το F1-Score είναι 0,822, υποδηλώνοντας ένα καλό συνολικό επίπεδο ακρίβειας.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

10 Συμπεράσματα και προτάσεις για μελλοντική εργασία

10.1 Συμπεράσματα

Συνοψίζοντας, στο πλαίσιο αυτής της έρευνας, η Σ. Σκορδίλη ανέλαβε στην εργασία της [Σ23] ένα σημαντικό καθήκον που περιλάμβανε την αυτοματοποιημένη συλλογή, αναγνώριση, επεξεργασία και κατηγοριοποίηση ειδησεογραφικών γεγονότων. Αυτή η προσπάθεια συνέβαλε σημαντικά στη βελτίωση της ποιότητας της πληροφορίας που αποθηκεύτηκε στη βάση δεδομένων μας. Από την πλευρά μου, η έρευνα επικεντρώθηκε στην αυτοματοποιημένη ανίχνευση επανεκδόσεων ειδησεογραφικών άρθρων και στην εύρεση ειδήσεων που περιγράφουν την εξέλιξη των γεγονότων με την πάροδο του χρόνου.

Στο πλαίσιο της παρούσας πτυχιακής εργασίας κατακτήθηκαν δύο κύριοι στόχοι. Πρώτον, να ανιχνευτούν επανεκδόσεις ειδησεογραφικών άρθρων, οι οποίες αποτελούν σημαντικό κομμάτι της βάσης δεδομένων μας. Αυτή η λειτουργικότητα βοηθά στη διατήρηση της ποιότητας και της εγκυρότητας των δεδομένων, αποτρέποντας τις διπλές εγγραφές και καθιστώντας τη βάση δεδομένων πιο οργανωμένη και διαχειρίσιμη.

Δεύτερον, ανιχνεύτηκαν ειδήσεις που περιγράφουν την εξέλιξη των γεγονότων με την πάροδο του χρόνου. Αυτό επιτρέπει στη βάση δεδομένων μας να παρέχει πολύτιμες πληροφορίες σχετικά με την ανάπτυξη των γεγονότων, τη σημαντικότητά τους, και τις σχέσεις μεταξύ τους στο πέρασμα του χρόνου.

Βασιζόμενοι στα αποτελέσματα της έρευνας, φαίνεται ότι το προτεινόμενο σύστημα εμφανίζει ικανοποιητική αυτοματοποιημένη απόδοση στους ζητούμενους στόχους. Ωστόσο, παραμένει χώρος για βελτιώσεις και προοπτικές για το μέλλον της υλοποίησης, τις οποίες θα εξετάσουμε στο ακόλουθο κεφάλαιο.

10.2 Προτάσεις για μελλοντική εργασία

10.2.1 Βελτίωση της ανακάλυψης follow-up ειδήσεων μέσω προηγμένων συστημάτων αναγνώρισης επώνυμων οντοτήτων (NER)

Ενώ η χρήση κεφαλαίων χρησίμευε παραδοσιακά ως βασικός δείκτης για την αναγνώριση των κατάλληλων ουσιαστικών και πιθανών follow-up ειδήσεων, η εμφάνιση προηγμένων συστημάτων αναγνώρισης ονομαστικών οντοτήτων (Named Entity Recognition - NER) αποτελεί μια πολλά υποσχόμενη μέθοδο για τον τρόπο αναγνώρισης και παρακολούθησης των εξελίξεων ειδήσεων. Αυτή η ενότητα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

περιγράφει πιθανές κατευθύνσεις και σκέψεις για τη εκμετάλλευση των συστημάτων NER για τη βελτίωση της ανακάλυψης follow-up ειδήσεων.

Οι μέθοδοι που βασίζονται στην κεφαλαιοποίηση βασίζονται κυρίως σε κειμενικά στοιχεία, συχνά παραβλέποντας τον πλούτο των πληροφοριών που περιέχονται σε εικόνες, βίντεο και άλλες μη κειμενικές πηγές. Τα μελλοντικά συστήματα εξαγωγής ειδήσεων θα μπορούσαν να ενσωματώσουν μοντέλα NER που είναι ικανά να αναλύουν και να εξάγουν οντότητες από διάφορες μορφές μέσων. Αυτό θα μπορούσε να περιλαμβάνει μοντέλα εκπαίδευσης σε ένα ποικίλο σύνολο δεδομένων που περιλαμβάνει τόσο κειμενικές όσο και οπτικές πληροφορίες, επιτρέποντας την αναγνώριση οντοτήτων όχι μόνο από κείμενο αλλά και από εικόνες, βίντεο και ηχητικά εφέ.

Οι παραδοσιακές προσεγγίσεις που βασίζονται στην κεφαλαιοποίηση είναι περιορισμένες στην ικανότητά τους να προσαρμόζονται σε ενημερώσεις ειδήσεων σε πραγματικό χρόνο και δυναμικές εξελίξεις. Το πιθανό πληροφοριακό σύστημα που βασίζεται στο NER θα μπορούσε να σχεδιαστεί για να παρέχει παρακολούθηση εξελισσόμενων ιστοριών σε πραγματικό χρόνο, παρακολουθώντας συνεχώς τις πηγές ειδήσεων και ενημερώνοντας ανάλογα τα μοντέλα αναγνώρισης οντοτήτων. Αυτό το σύστημα ενδέχεται να χρησιμοποιεί τεχνικές όπως η μηχανική μάθηση, όπου τα μοντέλα βελτιώνουν επαναληπτικά την κατανόησή τους για οντότητες με βάση την ανατροφοδότηση από νέα δεδομένα.

Η ενσωμάτωση προηγμένων συστημάτων αναγνώρισης επώνυμων οντοτήτων στην ανακάλυψη ειδήσεων που ακολουθεί έχει τη δυνατότητα να φέρει επανάσταση στον τρόπο με τον οποίο εντοπίζονται και παρακολουθούνται οι εξελίξεις των ειδήσεων. Αξιοποιώντας τις δυνατότητες των μοντέλων NER μπορούμε να δημιουργήσουμε πιο έξυπνο και εξατομικευμένο σύστημα που ανταποκρίνεται στο εξελισσόμενο τοπίο πληροφοριών.

10.2.1.1 Γεωαναφορά δεδομένων (Georeferenced data)

Τα συστήματα αναγνώρισης επώνυμων οντοτήτων (NER) μπορούν να συμβάλλουν σημαντικά στη δημιουργία γραφημάτων και οπτικοποιήσεων δεδομένων με γεωαναφορά, βοηθώντας στον εντοπισμό και την εξαγωγή πληροφοριών που σχετίζονται με την τοποθεσία από τα άρθρα ειδήσεων.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Κατά την επεξεργασία άρθρων ειδήσεων, αυτά τα συστήματα μπορούν να αναγνωρίσουν τοπωνύμια, γεωγραφικές συντεταγμένες και άλλες αναφορές τοποθεσίας, που χρησιμεύουν ως βάση για δεδομένα γεωαναφοράς. Μόλις εξαχθούν οι οντότητες τοποθεσίας, τα συστήματα NER μπορούν να τις συσχετίσουν αυτόματα με γεωγραφικές συντεταγμένες. Επιπλέον, μπορούν να εμπλουτίσουν δεδομένα γεωαναφοράς παρέχοντας πρόσθετο πλαίσιο σχετικά με τοποθεσίες. Για παράδειγμα, μπορούν να προσδιορίσουν εάν μια τοποθεσία είναι πόλη, πολιτεία, χώρα ή ορόσημο, κάτι που μπορεί να είναι χρήσιμο για την κατηγοριοποίηση και την οπτικοποίηση δεδομένων.

Με τα δεδομένα γεωαναφοράς που εξάγονται από τα συστήματα NER, μπορούν να δημιουργηθούν διάφοροι τύποι γραφημάτων και απεικονίσεων, όπως χάρτες θερμότητας, γραφήματα διασποράς και γραφήματα δικτύου για την αναπαράσταση της χωρικής κατανομής και των συνδέσεων συμβάντων ή σημείων δεδομένων. Τα γραφήματα δεδομένων με γεωαναφορά μπορούν να ενσωματωθούν σε διαδραστικά εργαλεία ή πλατφόρμες που επιτρέπουν στους χρήστες να εξερευνούν και να αλληλεπιδρούν με τα δεδομένα.

Συνδυάζοντας τα συστήματα NER με εργαλεία επεξεργασίας γεωχωρικών δεδομένων και βιβλιοθήκες οπτικοποίησης, είναι δυνατό να αυτοματοποιηθεί η εξαγωγή, η χαρτογράφηση και η οπτικοποίηση δεδομένων γεωαναφοράς από πηγές κειμένου, διευκολύνοντας την απόκτηση πληροφοριών από άρθρα ειδήσεων.

10.2.2 Χρήση Βαθιάς Μάθησης για την περαιτέρω αυτοματοποίηση του πληροφοριακού συστήματος

Η χρήση μέτρων ομοιότητας στην αναγνώριση διπλότυπων ειδήσεων αποτελεί σημαντικό στάδιο της διαδικασίας. Αυτά τα μέτρα μπορούν να περιλαμβάνουν την σύγκριση της σημασιολογίας των ειδήσεων, τον υπολογισμό της ομοιότητας μεταξύ των κειμένων ή την εφαρμογή τεχνικών όπως οι αναπαραστάσεις Word2Vec ή Doc2Vec. Παρόλα αυτά, αυτές οι προσεγγίσεις είναι περιορισμένες στον βαθμό που μπορούν να αντιμετωπίσουν περίπλοκα προβλήματα.

Η βαθιά μάθηση μπορεί να διαδραματίσει σημαντικό ρόλο στην αυτοματοποίηση ενός πληροφοριακού συστήματος που χρησιμοποιείται για την εύρεση διπλοτύπων ειδησεογραφικών γεγονότων. Οι αρχιτεκτονικές νευρωνικών δικτύων, όπως τα RNN δίκτυα, τα LSTM δίκτυα και οι αρχιτεκτονικές βασισμένες σε

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Transformers, δίνουν τη δυνατότητα στο σύστημα να αντιληφθεί την συντακτική και σημασιολογική δομή του κειμένου, καθιστώντας τη βαθιά μάθηση ιδιαίτερα αποτελεσματική για τον εντοπισμό διπλοτύπων ειδήσεων.

Τα νευρωνικά δίκτυα μπορούν να εξάγουν σύνθετες αναπαραστάσεις από τα κείμενα, εντοπίζοντας όχι μόνο κοινές λέξεις, αλλά και τη σημασιολογική σχέση μεταξύ τους. Αυτό σημαίνει ότι το σύστημα μπορεί να αντιληφθεί ακόμη και μεταφορές νοήματος, αναφορές σε παλαιότερες ειδήσεις, ή ακόμη και συγκείμενα που προϋπήρχαν της ειδήσεως.

Συνεπώς, η βαθιά μάθηση αντιπροσωπεύει μια ισχυρή προσέγγιση για τον εντοπισμό διπλοτύπων ειδήσεων, καθώς είναι ικανή να αντιμετωπίσει την πολυπλοκότητα της φυσικής γλώσσας και τη σημασιολογική ανάλυση περιεχομένου. Μπορεί να αυτοματοποιήσει σημαντικά τη διαδικασία εύρεσης διπλότυπων ειδησεογραφικών γεγονότων, 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 εκπαίδευση αυτών των μοντέλων απαιτεί σημαντικά δεδομένα με ετικέτα, υπολογιστικούς πόρους και εξειδίκευση σε τεχνικές βαθιάς μάθησης.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Παραρτήματα: Πηγαίος Κώδικας για τη Δημιουργία Γραφημάτων

Στο παράρτημα αυτό παρουσιάζονται οι πηγαίοι κώδικες που χρησιμοποιήθηκαν για τη δημιουργία των γραφημάτων που παρουσιάστηκαν στην πτυχιακή εργασία. Ο κώδικας είναι γραμμένος σε γλώσσα προγραμματισμού Python.

1 Γράφημα Word Cloud

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

# Load the Excel file
excel_file = 'ΠΡΩΤΟΤΥΠΑ_ΓΕΓΟΝΟΤΑ.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'
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 1: Γράφημα Word Cloud

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

2 Γράφημα Word Cloud ανά ημέρα

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

# Load the Excel file
excel_file = 'ΔΙΠΛΟΤΥΠΙΑ_ΓΕΓΟΝΟΤΑ.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_combined'
os.makedirs(output_dir, exist_ok=True)

# Create a DataFrame to hold combined data from all sheets
combined_df = pd.DataFrame(columns=['pubdate', 'title'])

# Process each sheet and combine data
for sheet_name in xlsx.sheet_names:

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

    # Append the data to the combined DataFrame
    combined_df = combined_df.append(df[['pubdate', 'title']], ignore_index=True)

# Convert the 'pubdate' column to a pandas datetime object
combined_df['pubdate'] = pd.to_datetime(combined_df['pubdate'])

# Group the data by day and concatenate titles within each day
day_groups = combined_df.groupby(combined_df['pubdate'].dt.date)['title'].apply(lambda x: ' '.join(x.astype(str)))

# Get the top 7 days with the most titles
top_7_days = day_groups.sort_values(ascending=False).index[:7]

# Iterate through each of the top 7 days and generate word clouds
for day in top_7_days:
    titles = day_groups[day]

    # 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 Top Day - {day}")
    plt.tight_layout()

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

    print(f"Word cloud saved for the top day: {day}")

print("Word clouds generated for the top 7 days.")
```

Εικόνα 2: Γράφημα Word Cloud ανά ημέρα

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

3 Γράφημα πλήθους ειδησεογραφικών γεγονότων ανά κατηγορία

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

file_path = 'ΠΡΩΤΟΤΥΠΙΑ_ΓΕΓΟΝΟΤΑ.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:
    # 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 3: Γράφημα πλήθους ειδησεογραφικών γεγονότων ανά κατηγορία

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

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.

[AAE12] Alsulami, B. S., Abulkhair, M. F., & Eassa, F. E. (2012). Near duplicate document detection survey. *International Journal of Computer Science and Communications Networks*, *2*(2), 147-151.

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

[APL98] Allan, J., Papka, R., & Lavrenko, V. (1998, August). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37-45).

[B23a] Bachmann Max (2023). python-Levenshtein 0.21.1. *Python Software Foundation*. Available at: https://pypi.org/project/python-Levenshtein/

[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

[CFGM02] Chowdhury, A., Frieder, O., Grossman, D., & McCabe, M. C. (2002). Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems (TOIS)*, *20*(2), 171-191.

[CSS02] Carthy, J., & Sherwood-Smith, M. (2002, October). Lexical chains for topic tracking. In *IEEE International Conference on Systems, Man and Cybernetics* (Vol. 7, pp. 5-pp). IEEE.

[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.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

[F08] Feng, A. (2008). *Incident threading in news*. University of Massachusetts Amherst.

[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).

[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/

[HYK10] Hajishirzi, H., Yih, W. T., & Kolcz, A. (2010, July). Adaptive near-duplicate detection via similarity learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 419-426).

[IH08] Ingram D., Henshall P. (2008) Available at: https://www.thenewsmanual.net/Manuals%20Volume%201/volume1_24.htm

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

[KR20] Kalbaliyev, E., & Rustamov, S. (2020, December). Text Similarity Detection using Machine Learning algorithms with Character-based similarity measures. In *Conference on multimedia, interaction, design and innovation* (pp. 11-19). Cham: Springer International Publishing.

[KRAA16] Khan, M., Rahman, A. U., Awan, M. D., & Alam, S. M. (2016, September). Normalizing digital news-stories for preservation. *In 2016 Eleventh International Conference on Digital Information Management (ICDIM)* (pp. 85-90). IEEE.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

[KUD18] Khan, M., Ur Rahman, A., & Daud Awan, M. (2018). Term-based approach for linking digital news stories. In *Digital Libraries and Multimedia Archives: 14th Italian Research Conference on Digital Libraries, IRCDL 2018, Udine, Italy, January 25-26, 2018, Proceedings 14* (pp. 127-138). Springer International Publishing.

[KUUN20] Khan, M., Ur Rahman, A., Ullah, M., & Naseem, R. (2020). The role of named entities in linking news articles during preservation. In *Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT'18), Vol. 1* (pp. 50-58). Springer International Publishing.

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

[LD07] LaFree, G., & Dugan, L. (2007). Introducing the global terrorism database. *Terrorism and political violence*, *19*(2), 181-204.

[M03] Makkonen, J. (2003). Investigations on event evolution on TDT. In *Proceedings of the HLT-NAACL 2003 student research workshop* (pp. 43-48).

[MGB14] Mahajan, R., Gupta, S. K., & Bedi, M. R. (2014). A survey of duplicate and near duplicate techniques. *International Journal of Engineering Research & Technology (IJERT)*, *5*.

[MPK20] Mishra, A. R., Panchal, V. K., & Kumar, P. (2020). Similarity Search based on Text Embedding Model for detection of Near Duplicates. *International Journal of Grid and Distributed Computing*, *13*(2), 1871-1881.

[NFPA04] Nallapati, R., Feng, A., Peng, F., & Allan, J. (2004, November). Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 446-453).

[NHD12] Nandagaonkar, S. S., Hanchate, D. B., & Deshmukh, S. N. (2012). Survey on event tracking and event evolution. *International Journal of Computer Technology and Applications*, *3*(1).

[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/

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

[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/

[PR] Patel, H., & Ravichandran, G. LEVENSHTEIN ALGORITHM.

[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

[R22] Reimers Nils (2022). sentence-transformers 2.2.2. *Python Software Foundation*. Available at: https://pypi.org/project/sentence-transformers/

[R23] Rose Erik (2023). more-itertools 10.1.0. *Python Software Foundation*. Available at: https://pypi.org/project/more-itertools/

[RB21] Rosvold, E. L., & Buhaug, H. (2021). GDIS, a global dataset of geocoded disaster locations. *Scientific data*, *8*(1), 61.

[RM23] Reimers N., Geigle G., et al. (2019). sentence-transformers 0.2.0. *Python Software Foundation.* Availabe at: https://pypi.org/project/sentence-transformers/0.2.0/

[S+12] Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E., & Williams, J. (2012). Social conflict in Africa: A new database. *International Interactions*, *38*(4), 503-511.

[S22] Sharma Shantanu (2022). FuzzyWuzzy Python library. *GeeksforGeeks.* Available at: https://www.geeksforgeeks.org/fuzzywuzzy-python-library/

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

[SCAD] "Climate Change and African Political Stability", Available at: https://www.strausscenter.org/ccaps-research-areas/social-conflict/database/

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

[SciD23] SciPy Developers (2023). scipy 1.11.2. *Python Software Foundation*. Available at: https://pypi.org/project/scipy/

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

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

[U09] Uyar, E. (2009). *Near-duplicate news detection using named entities* (Doctoral dissertation, Bilkent Universitesi (Turkey)).

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

[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.

[WC07] Wei, C. P., & Chang, Y. H. (2007). Discovering event evolution patterns from document sequences. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *37*(2), 273-283.

[WD20] Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, *11*(9), 421.

[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).

[X+11] Xiao, C., Wang, W., Lin, X., Yu, J. X., & Wang, G. (2011). Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems (TODS)*, *36*(3), 1-41.

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

[YPC98] Yang, Y., Pierce, T., & Carbonell, J. (1998, August). A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 28-36).

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

[YSW09] Yang, C. C., Shi, X., & Wei, C. P. (2009). Discovering event evolution graphs from news corpora. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(4), 850-863.

[Δ18] Δεβελάσκα Χ., (2018). Μετρώντας τη σημασιολογική απόσταση για τη δημιουργία ενός λεξικού στηριγμένου σε σώμα κειμένων δημοσιογραφικού λόγου. MSc Εφαρμοσμένης Γλωσσολογίας, Φιλοσοφική Σχολή Α.Π.Θ., Θεσσαλονίκη.

[Σ23] Σπυριδούλα Σκορδίλη: "Δημιουργία μιας ελληνικής βάσης δεδομένων παγκοσμίων γεγονότων: συλλογή, αναγνώριση και κατηγοριοποίηση γεγονότων σε πραγματικό χρόνο", Πτυχιακή Εργασία, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας, Οκτώβριος 2023.

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38

Institutional Repository - Library & Information Centre - University of Thessaly
02/05/2026 08:02:21 EEST - 216.73.217.38