ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Διπλωματική Εργασία

Διαδικτυακή εφαρμογή για την ανάλυση κειμένων

Web-based environment for text analysis

του

ΚΑΡΑΒΑΣΙΛΗ ΦΩΚΙΩΝ
MIS20006

Επιβλέπων Καθηγητής: Ευαγγελίδης Γεώργιος

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στα Πληροφοριακά Συστήματα

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον καθηγητή μου και επιβλέποντα της παρούσας διπλωματικής εργασίας κο Γεώργιο Ευαγγελίδη για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφερε καθ' όλη τη διάρκεια εκπόνησης της εργασίας μου, καθώς και για την εμπιστοσύνη που μου έδειξε όταν δέχτηκε να αναλάβει την επίβλεψη της.

Επιπλέον, οφείλω να ευχαριστήσω την οικογένεια μου για την αμέριστη συμπαράσταση και στήριξη που μου προσέφερε κατά τη διάρκεια φοίτησης μου. Όλοι μαζί αλλά και ο καθένας τους ξεχωριστά με βοήθησαν με τον τρόπο τους στην επιτυχή ολοκλήρωση της διπλωματικής μου εργασίας.

Περίληψη

Σε μια εποχή όπου η παραγωγή δεδομένων αυξάνεται και το κόστος συγκέντρωσης και συντήρησης τους μικραίνει, καθώς η τεχνολογία εξελίσσεται, ο τομέας της ανάλυσης κειμένων βρίσκει ολοένα και περισσότερες εφαρμογές. Ένα μεγάλο μέρος των δεδομένων που παράγονται σήμερα αποτελείται από κειμενικά δεδομένα, τα οποία αν και είναι εύκολο να συλλεχθούν, η μετέπειτα ανάλυση τους για την εξαγωγή χρήσιμων πληροφοριών αποτελεί μεγάλη πρόκληση. Η επιστημονική κοινότητα έχει κάνει σημαντικά βήματα και συνεχίζει να διευρύνει τις γνώσεις της πάνω σε αυτόν τον τομέα. Στην παρούσα εργασία γίνεται μια ανασκόπηση του θέματος. Θα γνωρίσουμε τα επιμέρους επιστημονικά πεδία που δραστηριοποιούνται στην ανάλυση κειμένων, θα δούμε μεθόδους και τεχνικές που λαμβάνουν χώρα κατά τη διαδικασία της ανάλυσης και θα επισημάνουμε ορισμένους τομείς εφαρμογής της. Καταλήγοντας, στα πλαίσια της εργασίας δημιουργήθηκε μια διαδικτυακή εφαρμογή ανάλυσης κειμένων (Texter) η οποία παρουσιάζεται εκτενώς στο 4ο κεφάλαιο της εργασίας.

Abstract

In an era where the production of data is increasing and the cost of gathering and maintaining is decreasing, as technology is developing, the field of text analysis is finding more and more applications. A considerable amount of the data produced today consists of textual data, which although easy to collect, their subsequent analysis to extract useful information has been proven a great challenge. The scientific community has made significant strides and continues to expand its knowledge in this field. In this paper, after a systematic literature search, a narrative review has been conducted. We will get to know the individual scientific fields which are involved in text analysis, we will see methods and techniques that take place during the analysis process, and we will highlight some areas of its application. Finally, for the needs of this thesis, an online application (Texter) for text analysis has been created, which is presented extensively in the 4th chapter.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1.Εισαγωγή

1.1 Περιγραφή του προβλήματος

Στη σημερινή εποχή παράγονται καθημερινά τεράστιες ποσότητες δεδομένων. Ο τύπος των δεδομένων που θα μας απασχολήσει στην εργασία αυτή είναι τα κειμενικά δεδομένα. Αν αναλογιστεί κανείς την τεράστια άνοδο του διαδικτύου και την αυξανόμενη χρήση των μέσων κοινωνικής δικτύωσης μπορεί εύκολα να συμπεράνει τον ανεξέλεγκτο ρυθμό παραγωγής κειμενικών δεδομένων. Σε μεγάλο ποσοστό τα δεδομένα αυτά είναι αδόμητα, ως εκ τούτου δεν μπορούν να χρησιμοποιηθούν για περαιτέρω επεξεργασία και εξαγωγή χρήσιμων πληροφοριών.

Αυτό το πρόβλημα λοιπόν καλείται να διαχειριστεί το πεδίο της εξόρυξης κειμένου. Πρόκειται για ένα πεδίο που λόγω των εξελίξεων στην τεχνολογία, σε υλικό και λογισμικό, κερδίζει ολοένα και περισσότερο έδαφος. Τα τελευταία χρόνια λόγω των αναγκών που έχουν προκύψει για διαχείριση, ταξινόμηση και επεξεργασία των κειμενικών δεδομένων, η εξόρυξη κειμένου έχει δικαίως κερδίσει την προσοχή της επιστημονικής κοινότητας. Υπολογίζεται ότι περίπου το 80% των πληροφοριών που έχουν στην κατοχή τους εταιρείες και οργανισμοί περιέχονται σε έγγραφα κειμένου (Akilan A., 2015).

Η εξόρυξη κειμένου χρησιμοποιεί νέες ερευνητικές μεθόδους και εργαλεία λογισμικού παρέχοντας σημαντική βοήθεια σε ερευνητές και επαγγελματίες διαφόρων κλάδων. Οι πολιτικές επιστήμες, η υγεία, οι κοινωνικές επιστήμες, το μάρκετινγκ είναι μόνο μερικοί από τους τομείς που μπορεί να βρει εφαρμογή και να υποστηρίξει η εξόρυξη κειμένου.

Ενώ στην περίπτωση των δομημένων δεδομένων αυτά είναι διαχειρίσιμα μέσω ενός συστήματος βάσεων δεδομένων, τα κειμενικά δεδομένα διαχειρίζονται με τη βοήθεια μηχανών αναζήτησης εξαιτίας της απουσίας δομής. Μια μηχανή αναζήτησης επιτρέπει στο χρήστη την εύρεση των ζητούμενων πληροφοριών μέσα από μια συλλογή εγγράφων με την υποβολή κάποιου ερωτήματος. Η αποτελεσματικότητα και η αποδοτικότητα της μηχανής αναζήτησης είναι ένα ζήτημα που απασχολεί κατά κύριο λόγο το πεδίο της ανάκτησης πληροφοριών, πεδίο το οποίο θα μπορούσαμε να πούμε ότι βρίσκεται υπό την ομπρέλα της εξόρυξης κειμένου. Ωστόσο, η έρευνα στην ανάκτηση πληροφοριών εστιάζει περισσότερο στη διευκόλυνση του χρήστη στην πρόσβαση πληροφοριών και όχι τόσο στην ανάλυση των πληροφοριών που έχει ως τελικό σκοπό την ανακάλυψη μοτίβων και συσχετίσεων, όπου είναι και η ουσία της εξόρυξης κειμένου. Ο στόχος της πρόσβασης στην πληροφορία είναι η διασύνδεση της αναζητούμενης πληροφορίας με τον εκάστοτε χρήστη την κατάλληλη στιγμή, δίνοντας λιγότερο έμφαση στην επεξεργασία και τον μετασχηματισμό των κειμενικών πληροφοριών. Η εξόρυξη κειμένου λοιπόν είναι αυτή που επιτυγχάνει να μας πάει ένα βήμα πιο μπροστά από την απλή πρόσβαση στην πληροφορία και μας βοηθά να αναλύσουμε και να επεξεργαστούμε πληροφορίες.

Βασιζόμενοι στη σημερινή τεχνολογία η συγκέντρωση πληροφοριών είναι το εύκολο κομμάτι, το δύσκολο κομμάτι είναι να βρούμε την κατάλληλη πληροφορία ανάλογα με τις απαιτήσεις

του χρήστη την κατάλληλη στιγμή. Καθώς το μέγεθος των κειμενικών συλλογών κλιμακώνεται εγείρεται η ανάγκη να προσδώσουμε κάποια δομή στα δεδομένα αυτά και ακολούθως να εξάγουμε συγκεκριμένα μοτίβα και χαρακτηριστικά που ανταποκρίνονται σε συγκεκριμένες πληροφοριακές ανάγκες.

1.2 Αντικείμενο και στόχοι της μελέτης

Η παρούσα εργασία χωρίζεται σε δύο σκέλη, το θεωρητικό και το πρακτικό. Αρχικά ξεκινώντας με το θεωρητικό κομμάτι, στόχος είναι η ευρύτερη κατανόηση του θέματος της ανάλυσης κειμένου έχοντας ως απώτερο σκοπό τη διερεύνηση του πεδίου της εξόρυξης κειμένου μέσα από την ανασκόπηση της βιβλιογραφίας. Η μελέτη αυτή αποσκοπεί να διευκολύνει τον αναγνώστη να κατανοήσει τι είναι η ανάλυση κειμένου, γιατί θεωρείται ένα από τα πιο ανερχόμενα επιστημονικά πεδία της εποχής μας και πως πραγματοποιείται βλέποντας τα διάφορα στάδια και τις τεχνικές που λαμβάνουν χώρα κατά τη διαδικασία της εξόρυξης κειμένου.

Όπως θα δούμε και στη συνέχεια της εργασίας κατά τη διαδικασία εξόρυξης κειμένου εμπλέκονται διάφορα ερευνητικά πεδία. Καθώς η πλήρης ανάλυση και ανάπτυξη της συμβολής αυτών των ερευνητικών πεδίων στην εξόρυξη κειμένου θα προκαλούσε μια σύγχυση στον αναγνώστη, αποφεύγονται οι πολλές τεχνικές λεπτομέρειες και γίνεται μια προσπάθεια να οριστεί η συμβολή αυτών των πεδίων στην εξόρυξη κειμένου μέσα από ένα πιο γενικό πλαίσιο.

Όσον αφορά το πρακτικό κομμάτι της εργασίας, δημιουργήθηκε μια διαδικτυακή εφαρμογή ανάλυσης κειμένου με στόχο να ενσωματώσει κάποια χαρακτηριστικά και τεχνικές που αναλύθηκαν κατά το θεωρητικό κομμάτι, δίνοντας έμφαση στην προεπεξεργασία των κειμενικών δεδομένων και στην εξαγωγή ποσοτικών συμπερασμάτων.

Καταλήγοντας, γίνεται μια σύντομη αναφορά σε εμπορικές και μη εφαρμογές ανάλυσης κειμένου όπου σε συνεργασία με το θεωρητικό περιεχόμενο της μελέτης αυτής γίνεται μια ανασκόπηση με στόχο την αυτοκριτική και την παροχή προτάσεων βελτίωσης.

1.3 Διάρθρωση της εργασίας

Η εργασία αποτελείται από πέντε (5) βασικές ενότητες:

Εισαγωγική Ενότητα: παρουσιάζεται το πρόβλημα που θα μελετηθεί και γίνεται μια εισαγωγή ώστε να κατανοήσει ο αναγνώστης το αντικείμενο και τους στόχους που πραγματεύεται η παρούσα εργασία, όπως επίσης αναλύεται και η δομή της εργασίας.

Ανάλυση Κειμένου: προσεγγίζεται το θέμα της ανάλυσης κειμένου από μια πιο φιλοσοφική σκοπιά και γίνεται μια ιστορική αναδρομή σχετικά με την εξέλιξη στον τομέα της ανάλυσης κειμένου.

Εξόρυξη Κειμένου - Text Mining: σε αυτή την ενότητα καταλήγουμε από το γενικότερο πλαίσιο της ανάλυσης κειμένου στο πιο ειδικό, αυτό της ανάλυσης κειμένου με τη βοήθεια υπολογιστικών συστημάτων. Παρουσιάζονται τα επιστημονικά πεδία που συνεισφέρουν στην εξόρυξη κειμένου καθώς και μέθοδοι και τεχνικές που χρησιμοποιούνται. Επίσης, γίνεται μια αναφορά σε εφαρμογές, εμπορικές και μη, τέτοιου τύπου.

Εφαρμογή Ανάλυσης Κειμένων Texter: αναφέρονται οι δυνατότητες και τα χαρακτηριστικά του Texter και επιπλέον παρέχεται το εγχειρίδιο χρήσης της εφαρμογής προσπαθώντας τόσο να καθοδηγήσει τον χρήστη για τη σωστή χρήση της εφαρμογής όσο και να τον βοηθήσει να κατανοήσει πως προκύπτουν και τι σημαίνουν τα αποτελέσματα της ανάλυσης.

Συμπεράσματα - Προτάσεις Βελτίωσης: είναι η καταληκτική ενότητα όπου παρουσιάζονται κάποια συμπεράσματα και προτάσεις μελλοντικής βελτίωσης για την εφαρμογή που αναπτύχθηκε στα πλαίσια της διπλωματικής αυτής εργασίας.

2.Τι είναι η ανάλυση κειμένου

2.1 Εισαγωγή

Η ανάλυση κειμένου μπορεί να σημαίνει διαφορετικά πράγματα ανάλογα με την επιστήμη που καλείται να ερμηνεύσει το τι είναι τελικά η ανάλυση κειμένου. Έτσι λοιπόν μπορεί να είναι η εκτενής μελέτη λογοτεχνικών κειμένων, η έρευνα και ο συλλογισμός των ιστορικών πηγών κάποιου κειμένου ή ακόμα και η υπολογιστική ανάλυση μεγάλων δεδομένων με τη χρήση τεχνικών εξόρυξης κειμένου. Μεταξύ αυτών των διαφορετικών προσεγγίσεων στην παρούσα εργασία εξετάζεται η ανάλυση κειμένου με τη βοήθεια υπολογιστικών συστημάτων. (https://www.ceu.edu/tanad/what)

2.2 Ορισμός

Η ανάλυση κειμένου είναι η διαδικασία ταξινόμησης και ανάλυσης δεδομένων που περιέχονται σε κάποιο κείμενο για ερευνητικούς σκοπούς. Η εξόρυξη κειμένου συνεπάγεται τον καθαρισμό, τη σήμανση, την οργάνωση και την ανάλυση του περιεχομένου μιας συλλογής εγγράφων. Με τη χρήση εργαλείων ψηφιακής ανάλυσης κειμένου, μπορούμε εύκολα να αναζητήσουμε και να εξετάσουμε συχνότητες λέξεων, μοτίβα και σχέσεις.
(https://guides.lib.fsu.edu/text-analysis/definitions)

2.3 Ιστορική Εξέλιξη

Για την καλύτερη κατανόηση της παρούσας κατάστασης των εργαλείων ανάλυσης κειμένου, θα γίνει μια σύντομη ανασκόπηση της ιστορίας σχετικά με τις πρακτικές και τα εργαλεία που χρησιμοποιήθηκαν και τη συνδρομή της τεχνολογίας σε θέματα ανάλυσης κειμένου.

Τα εργαλεία ανάλυσης κειμένου έχουν τις ρίζες τους στα έντυπα ευρετήρια όρων. Τα ευρετήρια όρων είναι ένα τυπικό ερευνητικό εργαλείο στις ανθρωπιστικές επιστήμες που φαίνεται να πρωτοεμφανίστηκε τον 13ο αιώνα. Σύμφωνα με τους Vannevar Bush και Douglas Engelbart πρόκειται για ένα εργαλείο που αυξάνει την επιστημονική μας εμβέλεια και βοηθά στην πνευματική εργασία.

Τα πρώτα εργαλεία ανάλυσης κειμένου που βασίζονται σε υπολογιστή σχεδιάστηκαν για να βοηθήσουν την κατασκευή εντύπων ευρετηρίων όρων. Ο πάτερ Roberto Busa στα τέλη της δεκαετίας του 1940 ήταν ένας από τους πρώτους που χρησιμοποίησε την τεχνολογία της πληροφορίας για την παραγωγή ευρετηρίων όρων με το ευρετήριο “Thomisticus”. Το έργο του αρχικά περιλάμβανε τη χρήση καρτών ευρετηρίου, στη συνέχεια πέρασε στην αναλογική τεχνολογία πληροφοριών τη δεκαετία του ‘50 για να ενσωματωθεί τελικά σε ηλεκτρονικούς υπολογιστές, μόλις αυτοί έγιναν διαθέσιμοι.

Τη δεκαετία του ‘60 και του ‘70 έγινε διαθέσιμη η πρώτη γενιά εργαλείων που προορίζονταν για χρήση από άλλους. Αυτά ήταν εργαλεία για κεντρικούς υπολογιστές που δεν απαιτούσαν οποιαδήποτε ανάμειξη από τον χρήστη για να λειτουργήσουν και σχεδιάστηκαν όπως τα εργαλεία του Μπούσα για να βοηθήσουν στην παραγωγή εντύπων ευρετηρίων όρων. Ουσιαστικά μέχρι εκείνη την περίοδο τα υπολογιστικά εργαλεία που είχαν δημιουργηθεί απευθύνονταν στους δημιουργούς των ευρετηρίων όρων. Κάποια από αυτά τα πρώιμα εργαλεία ήταν το COCOA (Count and Concordance generation on the Atlas) και το OCP(Oxford Concordance Program), με το δεύτερο να προκύπτει το 1978 όταν το υπολογιστικό τμήμα του πανεπιστημίου της Οξφόρδης ανέλαβε να βελτιώσει το COCOA.

Με τη διαθεσιμότητα και την αυξανόμενη δύναμη των μικροϋπολογιστών τη δεκαετία του ‘80 τα εργαλεία ανάλυσης κειμένου μεταπήδησαν από τους κεντρικούς υπολογιστές στους προσωπικούς υπολογιστές. Το OCP εξελίχθηκε σε Micro-OCP και κυκλοφόρησαν νέα προγράμματα, φτιαγμένα για προσωπικούς υπολογιστές, όπως το Brigham Young Concordance (BYC) που αργότερα μετονομάστηκε και κυκλοφόρησε στο εμπόριο με το όνομα WordCruncher και το περιβάλλον TACT που αναπτύχθηκε στο Πανεπιστήμιο του Τορόντο και κυκλοφόρησε το 1989, όπου και παρουσιάστηκε στο συνέδριο ACH/ ALLC. Με τα εργαλεία αυτά να γίνονται πλέον διαθέσιμα στους ερευνητές στον προσωπικό τους σταθμό εργασίας ο τρόπος με τον οποίο χρησιμοποιούνται άλλαξε ριζικά.

Πλέον ο μελετητής είχε άμεση πρόσβαση στα εργαλεία ανάλυσης κειμένου όποτε και όπου το χρειαζόταν. Στην πραγματικότητα αυτό σήμαινε πως ο ερευνητής δεν εξαρτιόταν πια από τα έντυπα ευρετήρια όρων όταν πραγματοποιούσε κάποια έρευνα καθώς είχε πλέον πρόσβαση σε ηλεκτρονικά εργαλεία. Αυτές οι αλλαγές στον τόπο και στο χρόνο που πραγματοποιείται μια μελέτη, οδήγησαν τους δημιουργούς εργαλείων ανάλυσης κειμένου στην κατασκευή πιο διαδραστικών μέσων, εκμεταλλευόμενοι το γεγονός πως οι ερευνητές έχουν άμεση πρόσβαση σε εργαλεία και ηλεκτρονικά κείμενα για μελέτη.

Καθώς τα διαθέσιμα εργαλεία ανάλυσης κειμένου αρχίζουν να αποκτούν ένα πιο διαδραστικό χαρακτήρα και η κοινότητα χρηστών μεγαλώνει και ωριμάζει είναι πλέον εμφανές πως μπορούν να τεθούν νέα ερωτήματα που τα έντυπα ευρετήρια όρων δε θα μπορούσαν να τα υποστηρίξουν. Από μια απλή αναζήτηση όρων φτάσαμε στο σημείο να μπορούμε να κάνουμε ερωτήματα σχετικά με περιστοιχιζόμενες λέξεις, αναζήτηση σύνθετων μοτίβων, μετρήσεις, σύγκριση λεξιλογίου και χαρακτήρων, οπτικοποίηση κειμένων και ακόμα περισσότερα. (Rockwell, Geoffrey, 2003)

2.4 Προσεγγίσεις της Ανάλυσης Κειμένων

Σύμφωνα με το βιβλίο των Gabe Ignatow και Rada Mihalcea, ‘An Introduction to Text Mining Research Design, Data Collection and Analysis’ το πεδίο της ανάλυσης κειμένων μπορεί να διαχωριστεί σε μια ποικιλία διαφορετικών προσεγγίσεων, η κάθε μια από τις οποίες βασίζεται στον διαφορετικό τρόπο εκτίμησης της χρήσης της γλώσσας. Οι έξι (6) πιο επικρατέστερες προσεγγίσεις παρουσιάζονται παρακάτω:

2.4.1 Ανάλυση συνομιλίας (Conversation Analysis)

Πρόκειται για την ανάλυση καθημερινών συνομιλιών όσων αφορά τον τρόπο με τον οποίο οι άνθρωποι αντιλαμβάνονται το νόημα της συνομιλίας στην οποία συμμετέχουν και το ευρύτερο πλαίσιο του οποίου η συνομιλία αποτελεί μέρος. Οι αναλυτές συνομιλίας δεν εστιάζουν μόνο σε αυτά που λέγονται στις καθημερινές συνομιλίες αλλά και στο πώς οι συμμετέχοντες χρησιμοποιούν πρακτικά τη γλώσσα για να ορίσουν τις καταστάσεις στις οποίες εμπλέκονται. (Gabe Ignatow, Rada Mihalcea 2018)

2.4.2 Ανάλυση Θέσεων Λόγου (Analysis of Discourse Positions)

Είναι μια προσέγγιση που επιτρέπει στους ερευνητές να ανακατασκευάσουν επικοινωνιακές αλληλεπιδράσεις μέσα από τις οποίες παράγονται κείμενα, αποκτώντας με αυτόν τον τρόπο καλύτερη κατανόηση της σημασίας τους από την οπτική του συγγραφέα. Οι λογικοί ρόλοι που υιοθετούν οι άνθρωποι στις καθημερινές επικοινωνιακές πρακτικές τους και η ανάλυση των τοποθετήσεων τους, επιτρέπει τη σύνδεση των κειμένων με τους κοινωνικούς χώρους από τους οποίους προέρχονται. (Gabe Ignatow, Rada Mihalcea 2018)

2.4.3 Κριτική Ανάλυση Λόγου (CDA - Critical Discourse Analysis)

Η κριτική ανάλυση λόγου βασίζεται στην έννοια της «διακειμενικότητας» του Fairclough (1995), η οποία είναι η ιδέα ότι ο γραπτός λόγος και η ομιλία των ανθρώπων επηρεάζονται από τα λεγόμενα του κοινωνικού τους περίγυρου. Πιο συγκεκριμένα είναι μια ποιοτική αναλυτική προσέγγιση για την κριτική περιγραφή, ερμηνεία και εξήγηση των τρόπων με τους οποίους ο λόγος κατασκευάζει, διατηρεί και δικαιολογεί τις κοινωνικές ανισότητες. (Mullet, 2018)

2.4.4 Ανάλυση Περιεχομένου (Content Analysis)

Η ανάλυση περιεχομένου υιοθετεί μια ποσοτική, επιστημονική προσέγγιση στην ανάλυση κειμένου. Σε αντίθεση με την κριτική ανάλυση λόγου (CDA), η ανάλυση περιεχομένου επικεντρώνεται στα ίδια τα κείμενα παρά στις σχέσεις των κειμένων με τα κοινωνικά και ιστορικά τους πλαίσια. Ένας από τους κλασικούς ορισμούς της ανάλυσης περιεχομένου την ορίζει ως «μια ερευνητική τεχνική με στόχο τη συστηματική - ποσοτική περιγραφή του έκδηλου περιεχομένου της επικοινωνίας» (Berelson, 1952, σ. 18). Σε πρακτικό επίπεδο, η ανάλυση περιεχομένου περιλαμβάνει την ανάπτυξη ενός πλαισίου κωδικοποίησης που εφαρμόζεται σε κειμενικά δεδομένα. Περιλαμβάνει κυρίως τη διάσπαση των κειμένων σε σχετικές ενότητες πληροφοριών προκειμένου να επιτραπεί η περαιτέρω κωδικοποίηση και κατηγοριοποίηση τους. (Gabe Ignatow, Rada Mihalcea 2018)

2.4.5 Foucauldian Analysis

Ο φιλόσοφος και ιστορικός Foucault (1973) ανέπτυξε εννοιολογικά τη διακειμενικότητα με τρόπο που διαφέρει σημαντικά από τη θεώρηση του Fairclough στην Κριτική Ανάλυση Λόγου

(CDA). Αντί να εντοπίζει κανείς την επιρροή εξωτερικών πηγών σε ένα κείμενο, για τον Foucault το νόημα ενός κειμένου προκύπτει σε σχέση με τις πηγές που εμπλέκονται σε αυτό. Στη διακειμενική ανάλυση του Foucault ο αναλυτής πρέπει να αναρωτηθεί για κάθε κείμενο ποιες είναι οι υποθέσεις που κάνει και με ποιες πηγές αλληλεπιδρά. Το νόημα του κειμένου επομένως προκύπτει από τις ομοιότητες και τις διαφορές αναφορικά με τα άλλα κείμενα και πηγές, και από τις έμμεσες υποθέσεις μέσα σε ένα κείμενο οι οποίες μπορούν να αναγνωριστούν με προσεκτική ανάγνωση. (Gabe Ignatow, Rada Mihalcea 2018)

2.4.6 Ανάλυση Κειμένων ως Κοινωνική Πληροφορία (Analysis of Texts as Social Information)

Αυτή η προσέγγιση αντιμετωπίζει τα κείμενα σαν αντανακλάσεις της πρακτικής γνώσης των συγγραφέων τους. Αυτός ο τύπος ανάλυσης είναι συνηθισμένος σε μελέτες θεμελιωμένης θεωρίας καθώς και σε εφαρμοσμένες μελέτες επιστημονικών πηγών. Το ενδιαφέρον για την ενημερωτική ανάλυση των κειμένων οφείλεται σε κάποιο βαθμό στην πρακτική τους αξία,
καθώς τα κείμενα που δημιουργούνται από τους χρήστες μπορούν να παρέχουν στους αναλυτές αξιόπιστες πληροφορίες σχετικά με την κοινωνική πραγματικότητα. Φυσικά, η ποιότητα των πληροφοριών για την κοινωνική πραγματικότητα που περιέχονται στα κείμενα, ποικίλλει ανάλογα με το επίπεδο γνώσεων του κάθε ατόμου που έχει συμμετάσχει στη δημιουργία του κειμένου. Ακολούθως οι πληροφορίες που παρέχονται μπορεί να είναι μεροληπτικές σε κάποιο βαθμό μιας και παρουσιάζονται από την ιδιαίτερη οπτική γωνία του συγγραφέα τους. (Gabe Ignatow, Rada Mihalcea 2018)

3.Εξόρυξη Κειμένου - Text Mining

3.1 Εισαγωγή

Όπως έγινε αντιληπτό από την προηγούμενη ενότητα, ενώ οι μέθοδοι ανάλυσης κειμένου έχουν μεγάλη ιστορία στις κοινωνικές επιστήμες (Roberts, 1997), η εξόρυξη κειμένου είναι σχετικά νέο διεπιστημονικό πεδίο που βασίζεται στην υπολογιστική επιστήμη. Τα τελευταία χρόνια η εξόρυξη κειμένου έχει αρχίσει να βρίσκει εφαρμογή και να δρα καταλυτικά ή και υποστηρικτικά σε πολλές επιστήμες όπως η ανθρωπολογία (Acerbi, Lampos, Garnett, & Bentley, 2013, Marwick, 2013), οι επικοινωνίες (Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015), τα οικονομικά (Levenberg, Pulman, Moilanen, Simpson, & Roberts, 2014), η εκπαίδευση (Evison, 2013), οι πολιτικές επιστήμες (Eshbaugh - Soha, 2010, Grimmer & Stewart, 2013), η ψυχολογία (Colley & Neal, 2012, Schmitt, 2005), η κοινωνιολογία (Bail, 2012, Heritage & Raymond, 2005, Mische, 2014) ακόμα και στον τομέα της υγείας (Raja U., Mitchell T., Day T., & Hardin J. 2008).

3.2 Ορισμός

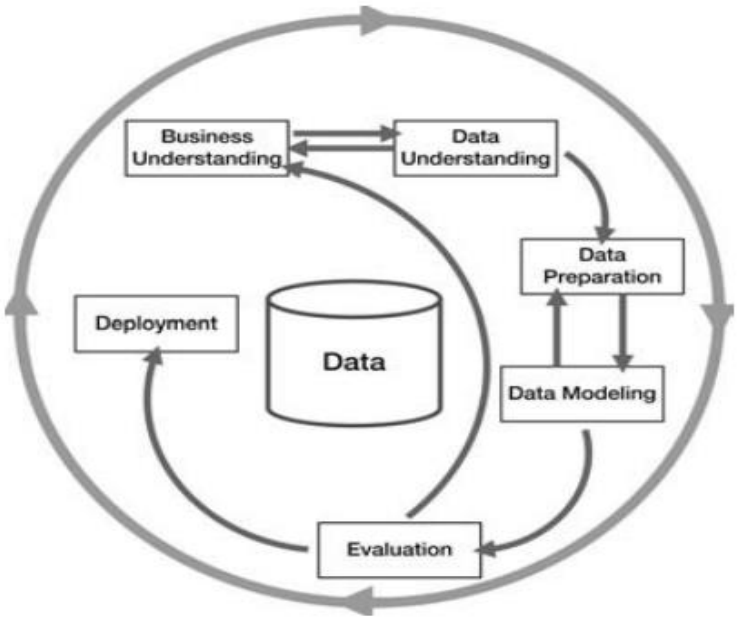
Η εξόρυξη κειμένου είναι η τέχνη της εξόρυξης δεδομένων από συλλογές δεδομένων κειμένου. Έχει ως στόχο την ανακάλυψη γνώσης (ή πληροφοριών, μοτίβων) από δεδομένα κειμένου, τα οποία είναι αδόμητα ή ημιδομημένα. Είναι ένα υποπεδίο της Εξόρυξης Δεδομένων (Data Mining - DM), το οποίο είναι επίσης γνωστό ως Ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD). Το KDD είναι η ανακάλυψη γνώσεων από διάφορες πηγές δεδομένων, συμπεριλαμβανομένων δεδομένων κειμένου, σχεσιακών βάσεων δεδομένων, δεδομένων Ιστού, δεδομένων καταγραφής χρηστών κ.λπ.

Η εξόρυξη κειμένου σχετίζεται επίσης με άλλα ερευνητικά πεδία, όπως η Μηχανική Μάθηση (Machine Learning - ML), η Ανάκτηση Πληροφοριών (Information Retrieval - IR), η Φυσική Επεξεργασία Γλώσσας (Natural Language Processing - NLP), η Εξαγωγή Πληροφοριών (Information Extraction - IE), η Στατιστική, η Αναγνώριση Προτύπων (Pattern Recognition - PR) και η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI). (Yanli Cai, Jian-Tao Sun, 2018)

3.3 Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases-KDD)

Η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (KDD) είναι η μη τετριμμένη διαδικασία προσδιορισμού έγκυρων, καινοτόμων, δυνητικά χρήσιμων και τελικά
κατανοητών μοτίβων στα δεδομένα. (Fayyad et al. 1996)

Η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (KDD) είναι μια διαδικασία που περιλαμβάνει αρκετά βήματα που εφαρμόζονται επαναληπτικά σε ένα σύνολο δεδομένων με σκοπό την εξαγωγή χρήσιμων μοτίβων. Κάποια από αυτά τα βήματα χρειάζονται ανατροφοδότηση (συμμετοχή) από τον χρήστη. Όπως ορίζεται από το μοντέλο CRoss Industry Standard Process for Data Mining (Crisp DM) τα κύρια βήματα είναι:
(1) επιχειρηματική κατανόηση,
(2) κατανόηση δεδομένων,
(3) προετοιμασία δεδομένων,
(4) μοντελοποίηση,
(5) αξιολόγηση,
(6) ανάπτυξη
Εκτός από το αρχικό πρόβλημα της ανάλυσης και κατανόησης της συνολικής διαδικασίας (δύο πρώτα βήματα) ένα από τα πιο χρονοβόρα βήματα είναι η προετοιμασία των δεδομένων. Το βήμα αυτό είναι ιδιαίτερης σημασίας για την εξόρυξη κειμένου καθώς κρίνεται σκόπιμο να εφαρμοστούν ειδικές τεχνικές προεπεξεργασίας για τη μετατροπή των δεδομένων κειμένου σε μορφή που θα είναι συμβατή με τους αλγόριθμους εξαγωγής δεδομένων (data mining algorithms). Στη συνέχεια της ενότητας θα γίνει εκτενέστερη αναφορά σε αυτές τις τεχνικές προεπεξεργασίας (Preprocessing Techniques).



Εικόνα 1: Phases of Crisp DM

3.4 Εξόρυξη Δεδομένων, Μηχανική Μάθηση και Στατιστική (Data Mining, Machine Learning and Statistics)

Καθώς η εξόρυξη δεδομένων και η ανακάλυψη γνώσης αποτελούν ακόμα ανοικτά ερευνητικά πεδία η επιστημονική κοινότητα δεν είναι ξεκάθαρη σχετικά με το πως ακριβώς ορίζονται. Από τη μια υπάρχει η θεώρηση πως η εξόρυξη δεδομένων είναι συνώνυμη της ανακάλυψης γνώσης, πράγμα που σημαίνει πως περιέχει όλες τις πτυχές της διαδικασίας ανακάλυψης γνώσης. Από την άλλη πλευρά υπάρχει η θεώρηση πως η εξόρυξη δεδομένων αποτελεί μέρος της διαδικασίας ανακάλυψης γνώσης και περιγράφει τη φάση της μοντελοποίησης, δηλαδή την εφαρμογή αλγορίθμων και μεθόδων για τον υπολογισμό και την αναζήτηση μοτίβων και μοντέλων. Μια ακόμα προσέγγιση είναι αυτή των Kumar & Joshi
(2003) όπου ορίζουν την εξόρυξη δεδομένων ως την αναζήτηση πολύτιμων πληροφοριών σε μεγάλες ποσότητες δεδομένων. (Hotho A., 2005)
Οι εξόρυξη δεδομένων έχει τις ρίζες της σε πολλά διαφορετικά ερευνητικά πεδία, γεγονός που τονίζει τον διεπιστημονικό χαρακτήρα αυτού του τομέα. Ακολουθεί μια σύντομη περιγραφή της σχέσης της εξόρυξης δεδομένων με τους τομείς των Βάσεων Δεδομένων, της Μηχανικής Μάθησης και της Στατιστικής.

Οι βάσεις δεδομένων είναι αναπόσπαστο κομμάτι στη διαδικασία ανάλυσης μεγάλων ποσοτήτων δεδομένων. Από αυτή την άποψη μια βάση δεδομένων δεν αντιπροσωπεύει μόνο το μέσο για συνεπή αποθήκευση και πρόσβαση στα δεδομένα, αλλά έχει ενεργό ρόλο στη διαδικασία εξόρυξης αφού η ανάλυση δεδομένων με τη χρήση αλγορίθμων εξόρυξης δεδομένων υποστηρίζεται από βάσεις δεδομένων. (Hotho A., 2005)

Η Μηχανική Μάθηση (ML) είναι ένας τομέας της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη τεχνικών που επιτρέπουν στους υπολογιστές να «μάθουν» μέσω της ανάλυσης ενός συνόλου δεδομένων. Η Μηχανική Μάθηση (ML) ασχολείται επίσης με την αλγοριθμική πολυπλοκότητα των υπολογιστικών υλοποιήσεων. (Hotho A., 2005)

Η στατιστική έχει τη βάση της στα μαθηματικά και ασχολείται με την ανάλυση εμπειρικών δεδομένων. Βασίζεται στη στατιστική θεωρία που είναι κλάδος των εφαρμοσμένων μαθηματικών. Μέσα στη στατιστική θεωρία, η τυχαιότητα και η αβεβαιότητα μοντελοποιούνται από τη θεωρία πιθανοτήτων. Σήμερα πολλές στατιστικές μέθοδοι χρησιμοποιούνται στον τομέα της Ανακάλυψης Γνώσης (KDD). (Hotho A., 2005)

3.5 Σχετικά Ερευνητικά Πεδία (Information Retrieval - IR, Natural Language Processing - NLP, Information Extraction - IE)

Όταν ορίσαμε την εξόρυξη κειμένου στην αρχή της ενότητας αναφερθήκαμε και σε κάποια ερευνητικά πεδία με τα οποία είναι άμεσα συσχετιζόμενη. Σε αυτήν την υποενότητα θα γνωρίσουμε εν συντομία κάποια από αυτά τα πεδία.

3.5.1 Ανάκτηση Πληροφοριών - Information Retrieval (IR)

Η ανάκτηση πληροφοριών είναι ο τρόπος να βρίσκεις έγγραφα που περιέχουν την απάντηση στην ερώτηση και όχι την απάντηση την ίδια (Hearst 1999). Για να πραγματοποιηθεί αυτό εφαρμόζονται στατιστικοί υπολογισμοί και μέθοδοι, για την αυτόματη επεξεργασία δεδομένων κειμένου, σε σύγκριση με το ερώτημα που καλούμαστε να απαντήσουμε. Η ανάκτηση πληροφοριών με την ευρύτερη έννοια ασχολείται με ολόκληρο το φάσμα της επεξεργασίας πληροφοριών, από την ανάκτηση δεδομένων έως την ανάκτηση γνώσης.

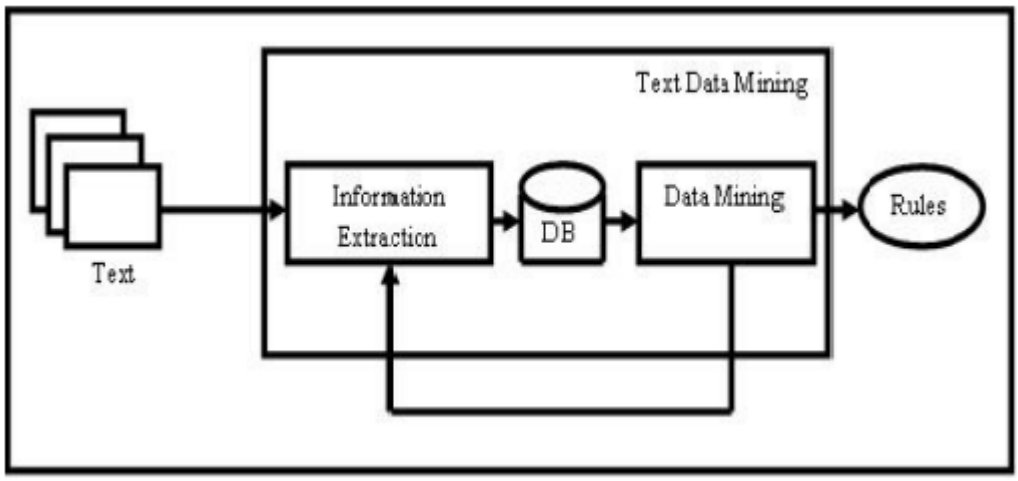
Η ανάκτηση πληροφοριών είναι σχετικά παλιά ερευνητική περιοχή η οποία εξαιτίας της ολοένα αυξανόμενης κειμενικής πληροφορίας έχει συνδράμει σε αρκετές εφαρμογές (Hotho A., 2005). Τέτοια παραδείγματα είναι συστήματα διαδικτυακών καταλόγων βιβλιοθήκης, συστήματα ηλεκτρονικής διαχείρισης αρχείων και το σχετικά πιο πρόσφατο παράδειγμα των διαδικτυακών μηχανών αναζήτησης που και αυτές με τη σειρά τους χρησιμοποιούν συστήματα ανάκτησης πληροφοριών (Mohan V., 2015).

3.5.2 Επεξεργασία Φυσικής Γλώσσας - Natural Language Processing (NLP)

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι ένας τομέας έρευνας που εξετάζει πως μπορούν να χρησιμοποιηθούν οι υπολογιστές για την κατανόηση και τον χειρισμό φυσικού γλωσσικού κειμένου. Οι ερευνητές αυτού του τομέα στοχεύουν στη συλλογή γνώσεων πάνω στο πως τα ανθρώπινα όντα κατανοούν και χρησιμοποιούν τη γλώσσα, έτσι ώστε να αναπτύξουν εργαλεία και τεχνικές που βοηθούν τα υπολογιστικά συστήματα να χειριστούν και να κατανοήσουν τη φυσική γλώσσα με σκοπό να εκτελέσουν τις ζητούμενες διεργασίες.
Οι αρχές της επεξεργασίας φυσικής γλώσσας βασίζονται σε μια πληθώρα άλλων επιστημονικών κλάδων όπως η επιστήμη των υπολογιστών και πληροφοριών, η γλωσσολογία, τα μαθηματικά, η ηλεκτρολογία και ηλεκτρονική μηχανική, η τεχνητή νοημοσύνη και ρομποτική, η ψυχολογία κ.α. Όσον αφορά τα ερευνητικά πεδία στα οποία η επεξεργασία φυσικής γλώσσας βρίσκει εφαρμογή, κάποια από αυτά είναι: η αυτόματη μετάφραση, οι διεπαφές χρηστών, η πολυγλωσσική και διαγλωσσική ανάκτηση πληροφοριών (multilingual and cross language information retrieval - CLIR), η αναγνώριση ομιλίας, η τεχνητή νοημοσύνη, τα ευφυή συστήματα κ.α. (S.Jusoh and H.M. Alfawareh, 2007)

3.5.3 Εξαγωγή Πληροφοριών - Information Extraction (IE)

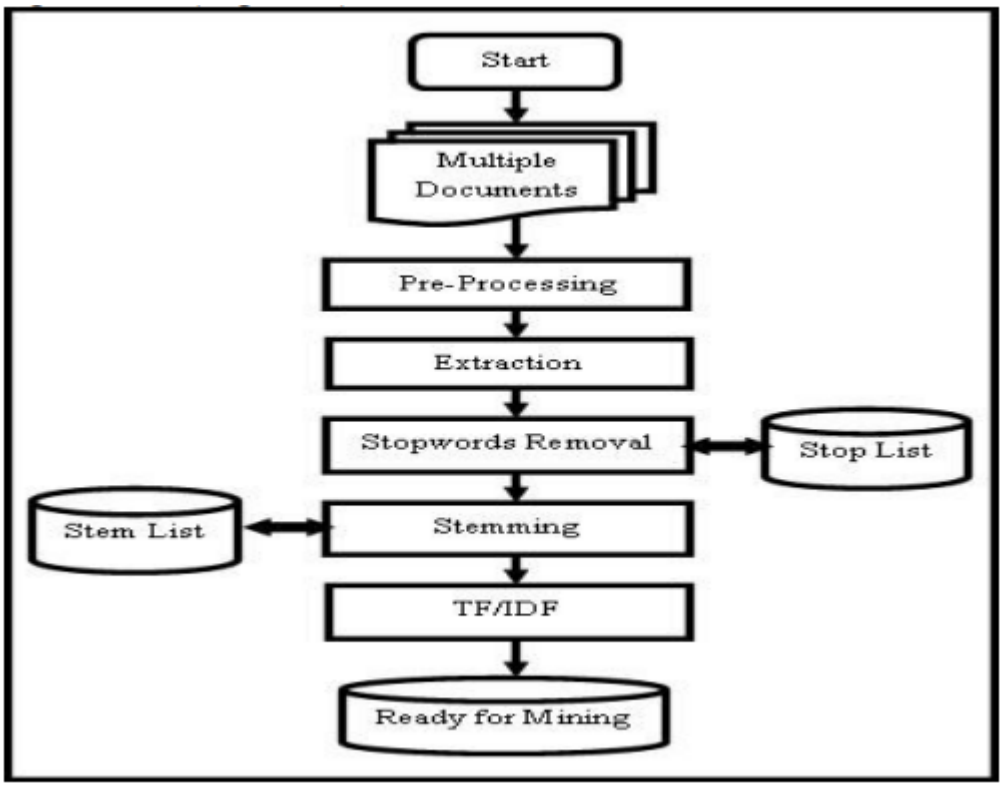
Η μέθοδος της εξαγωγής πληροφοριών εντοπίζει λέξεις κλειδιά και σχέσεις μεταξύ των κειμένων. Αυτό το πετυχαίνει αναζητώντας προκαθορισμένες ακολουθίες σε ένα κείμενο, μια διαδικασία που ονομάζεται αντιστοίχιση μοτίβων (pattern matching). Το λογισμικό εντοπίζει τις σχέσεις μεταξύ όλων των προσδιορισμένων στοιχείων με στόχο να παρέχει στο χρήστη ουσιαστικές πληροφορίες. Αυτή η τεχνολογία είναι ιδιαίτερα χρήσιμη όταν έχουμε να κάνουμε με μεγάλους όγκους κειμενικών δεδομένων. Παραδοσιακά η εξόρυξη δεδομένων υποθέτει ότι οι πληροφορίες που θα ‘εξορυχθούν’ έχουν ήδη τη μορφή μιας σχεσιακής βάσης δεδομένων. Δυστυχώς όμως σε πολλές περιπτώσεις οι ηλεκτρονικές πληροφορίες είναι απλώς διαθέσιμες σε έγγραφα φυσικής γλώσσας και όχι δομημένες σε βάσεις δεδομένων (Dr. S. Vijayarani et al., 2015). Η διαδικασία εξαγωγής κειμένου φαίνεται στο παρακάτω σχήμα της εικόνας 2.



Εικόνα 2:Process of Text Extraction

3.6 Μέθοδοι Προεπεξεργασίας στην Εξόρυξη Κειμένου (Preprocessing Techniques for Text Mining)

Η προεπεξεργασία είναι μείζονος σημασίας κατά τη διαδικασία της εξόρυξης κειμένου. Ειδικά όταν πρόκειται για μεγάλες συλλογές εγγράφων κρίνεται σκόπιμο να προεπεξεργαστούμε τα έγγραφα κειμένου και να αποθηκεύσουμε τις πληροφορίες σε μια δομή δεδομένων, η οποία θα είναι κατάλληλη για περαιτέρω επεξεργασία. Αποτελεί το πρώτο βήμα κατά τη διαδικασία εξόρυξης κειμένου και παρακάτω θα δούμε αναλυτικότερα τα βήματα κλειδιά της προεπεξεργασίας κειμένου: εξαγωγή (extraction), απομάκρυνση στοπ λέξεων (stop words removal), αποκοπή καταλήξεων (stemming), TF/IDF αλγόριθμοι (Term Frequency–Inverse Document Frequency algorithms). (βλέπε εικόνα 3)



Εικόνα 3: Στάδια Προεπεξεργασίας στην Εξόρυξη Κειμένου

3.6.1 Εξόρυξη - Extraction

Για να πάρουμε όλες τις λέξεις που χρησιμοποιούνται σε ένα δεδομένο κείμενο, χρειάζεται αρχικά να ακολουθήσουμε τη διαδικασία της διακριτικοποίησης (tokenization) κατά την οποία ένα έγγραφο κειμένου χωρίζεται σε λεξικογραφικές μονάδες (tokens), ενώ παράλληλα αφαιρούνται όλα τα σημεία στίξης και κάθε μη κειμενικός χαρακτήρας αντικαθίσταται με κενό. Η μορφή που θα πάρει το έγγραφο μετά την επεξεργασία που προαναφέρθηκε είναι καταλληλότερη για περαιτέρω επεξεργασία. Αυτή η διαδικασία θα εφαρμοστεί σε κάθε έγγραφο που ανήκει σε μια συλλογή και έτσι θα προκύψει ένα σύνολο διαφορετικών λέξεων

που συγκεντρώθηκαν από τη συγχώνευση των εγγράφων μας. Το σύνολό αυτό ονομάζεται λεξικό της συλλογής εγγράφων. (Hotho A., 2005)

3.6.2 Απομάκρυνση Στοπ Λέξεων - Stop Words Removal

Οι 'stop words' αποτελούν κομμάτι της φυσικής γλώσσας. Ο λόγος που αυτές οι λέξεις πρέπει να αφαιρεθούν από το κείμενο είναι διότι κάνουν το κείμενο να φαίνεται ανούσια φορτωμένο με λέξεις που είναι ουσιαστικά άχρηστες για τους αναλυτές. Λέξεις όπως άρθρα, σύνδεσμοι, προθέσεις κ.λπ. μας παρέχουν ελάχιστες ή καθόλου πληροφορίες σχετικά με το περιεχόμενο των κειμένων. Επιπλέον, λέξεις που εμφανίζονται αρκετά συχνά είναι πιθανό να δώσουν ελάχιστες πληροφορίες για το περιεχόμενο και τον διαχωρισμό των εγγράφων, όπως επίσης και λέξεις που εμφανίζονται πολύ σπάνια μπορεί να μην τυγχάνουν ιδιαίτερης στατιστικής σημασίας, επομένως θα ήταν χρήσιμο τέτοιες λέξεις να απομακρύνονται από το λεξικό (Frakes & Baeza-Yates 1992, Dr. S. Vijayarani, 2015). Ακολουθούν τέσσερις μέθοδοι που χρησιμοποιούνται για την απομάκρυνση στοπ λέξεων:

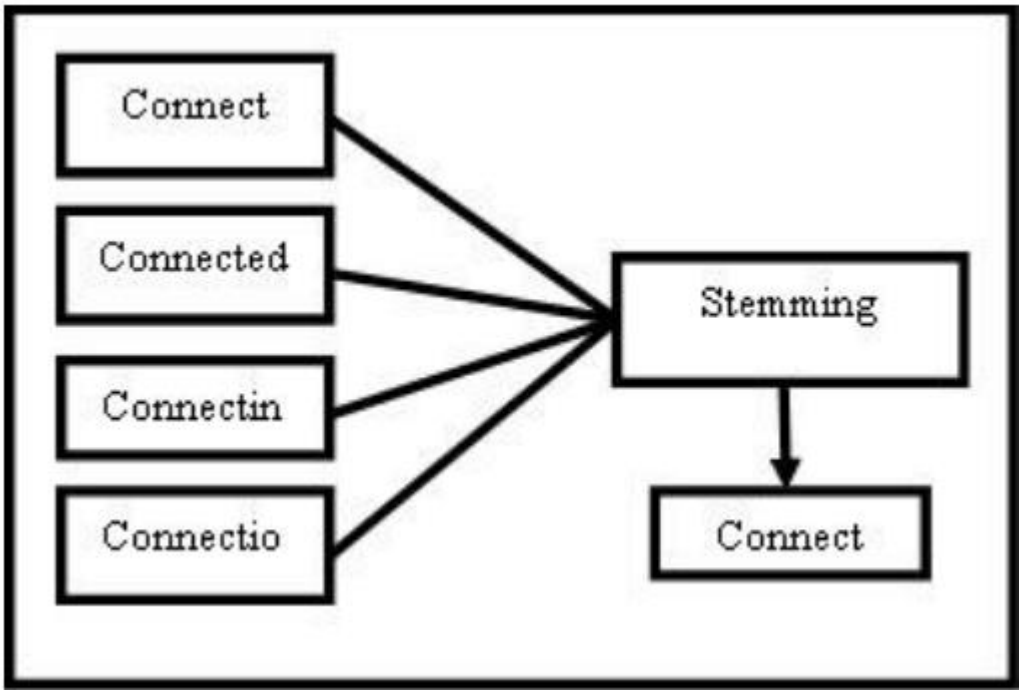
1. Η κλασσική μέθοδος: βασίζεται στην απομάκρυνση λέξεων που περιέχονται σε προσυμπληρωμένες λίστες (Anjali Ganesh Jivani, 2011).
2. Μέθοδος βασισμένη στο νόμο του Zipf (Z-Methods): κατά την οποία απομακρύνονται οι πιο συχνά εμφανιζόμενες λέξεις και οι λέξεις που εμφανίζονται μόνο μια φορά. Επίσης, εξετάζεται η απομάκρυνση των λέξεων με χαμηλό δείκτη IDF (Inverse Document Frequency). Περισσότερες πληροφορίες για τον IDF θα δοθούν στη συνέχεια της υποενότητας (Anjali Ganesh Jivani, 2011, Deepika Sharma, 2012).
3. Η μέθοδος αμοιβαίας πληροφόρησης (Mutual Information Method - MI): είναι μια μέθοδος που υπολογίζει τις αμοιβαίες πληροφορίες μεταξύ ενός δεδομένου όρου και μιας κλάσης εγγράφων, υποδεικνύοντας το πόσο σχετικές είναι οι πληροφορίες που μας δίνει ο όρος σε σχέση με τη δοθείσα κλάση. Επομένως, ένα χαμηλό σκορ θα σήμαινε ότι ο όρος πρέπει να απομακρυνθεί (Anjali Ganesh Jivani, 2011, Deepika Sharma, 2012).
4. Η μέθοδος της τυχαίας δειγματοληψίας βάσει όρου (Term Based Random Sampling - TBRS): Αυτή η μέθοδος λειτουργεί επαναληπτικά σε ξεχωριστά κομμάτια δεδομένων που επιλέγονται τυχαία. Κατατάσσει τους όρους σε κάθε κομμάτι με βάση τις τιμές που προκύπτουν χρησιμοποιώντας το μέτρο απόκλισης των Kullback-Leibler όπως φαίνεται στην εξίσωση 1.

   $$dx\ (t) = Px\ (t).log2\frac{Px\ (t)}{p\ (t)} \quad (1)$$

   Όπου Px (t) είναι η κανονικοποιημένη συχνότητα ενός όρου t εντός μιας μάζας x, και P(t) είναι η κανονικοποιημένη συχνότητα του όρου t σε ολόκληρη τη συλλογή. Η τελική λίστα με τις λέξεις που χρήζουν απομάκρυνσης δημιουργείται λαμβάνοντας τους λιγότερο σχετικούς όρους από όλα τα κομμάτια, και αφαιρώντας τις πιθανές επικαλύψεις (Anjali Ganesh Jivani, 2011).

3.6.3 Αποκοπή Καταλήξεων - Stemming

Αυτή η μέθοδος χρησιμοποιείται για να προσδιορίσει τη ρίζα/κορμό των λέξεων που αποτελούν το κείμενο. Σκοπός είναι να αφαιρεθούν τα διάφορα επιθήματα για να μειωθεί ο αριθμός των λέξεων και να εξοικονομήσουμε χρόνο και χώρο μνήμης. Ουσιαστικά μια ρίζα αντιπροσωπεύει μια ομάδα λέξεων με ίδια ή κοντινή σημασία, επομένως στο τέλος της διαδικασίας κάθε λέξη θα αντιπροσωπεύεται από τη ρίζα (stem) της. Μπορεί να γίνει πιο εύκολα κατανοητό κοιτώντας το παράδειγμα της εικόνας 4.



Εικόνα 4:Stemming Process

Κατά τη διαδικασία αποκοπής καταλήξεων η μορφολογική μετάβαση των λέξεων από την αρχική τους μορφή στη ρίζα τους πρέπει να γίνεται με την προϋπόθεση ότι κάθε μια από αυτές είναι σημασιολογικά συναφής. Υπάρχουν δύο (2) κανόνες που πρέπει να τηρούνται κατά τη διαδικασία του stemming:

- Λέξεις που δεν έχουν την ίδια σημασία πρέπει να διατηρούνται χωριστά
- Λέξεις που είναι μορφολογικά και σημασιολογικά παρόμοιες θα πρέπει να συνοψίζονται στην ίδια ρίζα (βλέπε εικόνα 4).

(Dr. S. Vijayarani, 2015, Hotho A. , 2005)

3.6.4 Αλγόριθμοι Αποκοπής Καταλήξεων - Stemming Algorithms

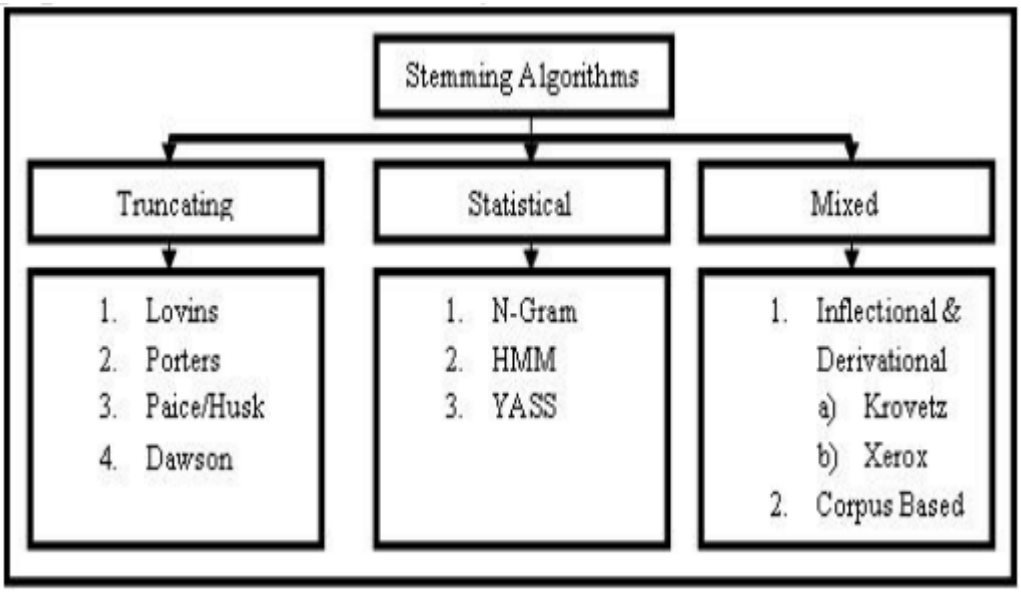
Οι αλγόριθμοι αποκοπής καταλήξεων μπορούν να ταξινομηθούν σε τρεις (3) κατηγορίες:

- Μέθοδοι περικοπής (truncating methods): όπως δηλώνει και το όνομα, αυτή η μέθοδος σχετίζεται με την αφαίρεση των προθημάτων ή των επιθημάτων των λέξεων. Η πιο κλασική εφαρμογή αυτής της μεθόδου αφαιρεί από τις λέξεις ένα συγκεκριμένο αριθμό (v) γραμμάτων, έχοντας ως βασικό μειονέκτημα ότι οι λέξεις με αριθμό γραμμάτων

μικρότερο ή ίσο με (v) παραμένουν ως έχουν. Μια ακόμα απλή προσέγγιση είναι ο αλγόριθμος της Donna Harman που συγχωνεύει τον ενικό και τον πληθυντικό αριθμό των ουσιαστικών στην αγγλική γλώσσα αφαιρώντας τις καταλήξεις των λέξεων που είναι στον πληθυντικό αριθμό.

- Στατιστικές Μέθοδοι (Statistical Methods): όπως και η προηγούμενη κατηγορία έτσι και αυτή αφαιρεί τα προσφύματα των λέξεων αφού πρώτα εφαρμοστούν κάποιες στατιστικές διαδικασίες.
- Μικτές μέθοδοι (Mixed Methods): αυτές οι μέθοδοι επιχειρούν να απομονώσουν τη ρίζα της λέξης στα κλιτά μέρη του λόγου όπως επίσης και τις ρίζες από παράγωγες λέξεις.

όπου κάθε κατηγορία έχει το δικό της χαρακτηριστικό τρόπο εύρεσης της ρίζας (stem) των παραλλαγών των λέξεων. Παρακάτω στην εικόνα 5 φαίνεται η ταξινόμηση των αλγορίθμων. (Deepika Sharma, 2012, Ms. Anjali Ganesh Jivani, 2011)



Εικόνα 5: Stemming Algorithms

3.6.5 TF/IDF Αλγόριθμοι (Term Frequency–Inverse Document Frequency Algorithms)

Η Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου (TF - IDF) είναι ένας στατιστικός δείκτης που μας δείχνει πόσο σημαντική είναι μια λέξη για ένα έγγραφο που ανήκει σε μια συλλογή εγγράφων. Χρησιμοποιείται συχνά ως παράγοντας στάθμισης κατά τη διαδικασία της ανάκτησης πληροφοριών και της εξόρυξης κειμένου. Η τιμή του δείκτη αυξάνεται αναλογικά με τον αριθμό των φορών που μια λέξη Α εμφανίζεται σε ένα έγγραφο, αλλά αντισταθμίζεται από τη συχνότητα της λέξης στη συλλογή εγγράφων. Αυτό μας βοηθά να διαχειριστούμε το γεγονός πως κάποιες λέξεις συναντιούνται πιο συχνά από κάποιες άλλες. Τέτοιου τύπου αλγόριθμοι μπορούν να χρησιμοποιηθούν στο φιλτράρισμα των στοπ λέξεων (stop words filtering) καθώς και κατά τη διαδικασία σύνοψης και ταξινόμησης κειμένων (text summarization/text classification) - σε επόμενη υποενότητα θα αναφερθούμε εκτενέστερα σε

αυτές τις 2 μεθόδους. Ο δείκτης TF/IDF προκύπτει από τη συγχώνευση του δείκτη Συχνότητας Όρου - TF και του δείκτη Αντίστροφης Συχνότητας Εγγράφου - IDF. Για να γίνει πιο κατανοητό, ο δείκτης TF ορίζεται ως ο συνολικός αριθμός των φορών που μια λέξη εμφανίζεται σε ένα έγγραφο, ενώ ο δείκτης IDF μπορεί να οριστεί σαν ένας συντελεστής βαρύτητας που μας δείχνει τη σημαντικότητα ενός όρου σε μια συλλογή εγγράφων.

$$\text{Tf}(t, d) = 0.5 + \frac{0.5 * f(t,d)}{Maximum\ Occurrences\ of\ Word} \quad (1)$$

$$\text{IDF}(t, d)\ \log = \frac{|D|}{(no.of\ doc., term\ t\ appears)} \quad (2)$$

*|D| : total number of documents in the corpus

$$\text{tfidf}(t, f, d) = \text{tf}(t, d) * \text{idf}(t, d) \quad (3)$$

Στις εξισώσεις (1) και (2) τα ft, d ορίζουν τη συχνότητα εμφάνισης ενός όρου t σε ένα έγγραφο d. Στην εξίσωση (3) ο δείκτης TF - IDF υπολογίζεται για κάθε όρο στο έγγραφο με τη χρήση των δεικτών TF (Tft, d) και IDF (idft, d). (Menaka S and Radha N, December 2013, S. Charanyaa and K. Sangeetha, February 2014)

3.7 Μέθοδοι και Τεχνικές Εξόρυξης Κειμένου (Text Mining Methods and Techniques)

3.7.1 Μοντέλα που χρησιμοποιούνται στην εξόρυξη κειμένου

Υπάρχει μια πληθώρα τεχνικών που έχουν αναπτυχθεί για την επίλυση του προβλήματος της εξόρυξης κειμένου. Αυτό το πρόβλημα δεν είναι άλλο από την ανάκτηση πληροφοριών σχετικά με τις απαιτήσεις του εκάστοτε χρήστη. Σύμφωνα με την ανάκτηση πληροφοριών υπάρχουν τέσσερις (4) βασικές μέθοδοι που χρησιμοποιούνται:

3.7.1.1 Term Based Method (TBM) - Μέθοδος βασισμένη στους όρους

Κάθε όρος σε ένα έγγραφο λογίζεται ως λέξη που έχει τη δική της σημασιολογική έννοια. Στη συγκεκριμένη μέθοδο το έγγραφο αναλύεται με βάση τους όρους που περιέχει και βασίζεται στην αυξημένη υπολογιστική απόδοση καθώς και στις θεωρίες για τη στάθμιση όρων. Αυτές οι μέθοδοι αναπτύχθηκαν τις τελευταίες δεκαετίες από τις κοινότητες της ανάκτησης πληροφοριών και της μηχανικής μάθησης. Πρόκληση αποτελούν τα προβλήματα της πολυσημίας και της συνωνυμίας. Κατά την πολυσημία από μια λέξη πηγάζουν πολλαπλά σημαινόμενα, ενώ κατά τη συνωνυμία πολλές λέξεις έχουν την ίδια ή σχεδόν την ίδια σημασία. Το πεδίο της ανάκτησης πληροφοριών (information retrieval) επιχειρεί να ξεπεράσει αυτή την πρόκληση παρέχοντας κάποιες μεθόδους βασισμένες στους όρους (term-based methods). (G. Salton and C. Buckley,1988)

3.7.1.2 Phrase Based Method (PBM) - Μέθοδος βασισμένη στις φράσεις

Οι φράσεις παρέχουν περισσότερες σημασιολογικές πληροφορίες και είναι λιγότερο ασαφής. Σε αυτή τη μέθοδο το έγγραφο αναλύεται με βάση τις φράσεις που περιέχει καθώς αυτές είναι λιγότερο διφορούμενες και πιο εύκολα διαφοροποιήσιμες από ότι οι μεμονωμένοι όροι. Πιθανές προκλήσεις της μεθόδου είναι:

- Οι φράσεις έχουν κατώτερες στατιστικές ιδιότητες σε σχέση με τους όρους,
- Έχουν χαμηλή συχνότητα εμφάνισης,
- Σε ένα έγγραφο συναντάται μεγάλος αριθμός περιττών φράσεων

(H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, 1998)

3.7.1.3 Concept Based Method (CBM) - Μέθοδος βασισμένη στις έννοιες

Σε αυτή τη μέθοδο οι όροι αναλύονται σε επίπεδο πρότασης και σε επίπεδο εγγράφου. Οι τεχνικές εξόρυξης κειμένου βασίζονται κυρίως στη στατιστική ανάλυση λέξεων και φράσεων. Η στατιστική ανάλυση της συχνότητας εμφάνισης των όρων σε ένα έγγραφο αποτυπώνει τη σημαντικότητα της λέξης ανεξαρτήτως του εγγράφου. Δύο (2) όροι μπορούν να έχουν την ίδια συχνότητα εμφάνισης στο ίδιο έγγραφο παρόλα αυτά ο ένας εξ αυτών μπορεί να συμβάλλει περισσότερο στην κατανόηση του περιεχομένου από ότι ο άλλος. Είναι λογικό λοιπόν να πρέπει να δοθεί περισσότερη προσοχή στον όρο με τη μεγαλύτερη συμβολή. Με τη συγκεκριμένη μέθοδο λοιπόν μπορεί να γίνει αποτελεσματική διάκριση μεταξύ των μη σημαντικών όρων και αυτών που έχουν μεγαλύτερη βαρύτητα και πιθανότατα να περιγράφουν το νόημα μιας πρότασης. Το εννοιολογικό αυτό μοντέλο βασίζεται σε τεχνικές επεξεργασίας φυσικής γλώσσας. (S. Shehata, F. Karray, and M. Kamel, 2006/2007)

3.7.1.4 Pattern Taxonomy Method (PTM) - Μέθοδος ταξινόμησης προτύπων

Στη μέθοδο ταξινόμησης προτύπων τα έγγραφα αναλύονται με βάση κάποια μοτίβα. Η ταξινόμηση πραγματοποιείται με τη χρήση σχέσεων “έχει ένα”, ”είναι ένα” (has - a, is - a relationship). Ο εντοπισμός προτύπων έχει μελετηθεί εκτενέστερα από τον κλάδο της εξόρυξης δεδομένων εδώ και πολλά χρόνια, άλλωστε τα μοτίβα μπορούν να ανακαλυφθούν με τη χρήση τεχνικών εξόρυξης δεδομένων (association rule mining, frequent item set mining, sequential pattern mining, closed pattern mining). Έρευνες έχουν δείξει πως η απόδοση αυτής της μεθόδου υπερτερεί από τις προηγούμενες τρεις (3). (Vijay Gaikwad, S., Chaugule, A., & Patil, P., 2014)

3.7.2 Μέθοδοι Εξόρυξης Δεδομένων για Κείμενα

Κατά κύριο λόγο η χρήση μεθόδων εξόρυξης δεδομένων σε συλλογές εγγράφων κειμένου έχουν ως στόχο να προσδώσουν μια δομή σε αυτές τις συλλογές. Μια δομημένη συλλογή εγγράφων μπορεί να διευκολύνει και να απλοποιήσει την πρόσβαση σε έγγραφα και πληροφορίες από τους χρήστες. Οι υπάρχουσες μέθοδοι για την οργάνωση συλλογών εγγράφων είναι είτε με την ανάθεση λέξεων κλειδιών που αντιστοιχίζονται με τα ανάλογα έγγραφα με βάση ένα δεδομένο σύνολο λέξεων κλειδιών (μέθοδοι ταξινόμησης - classification ή κατηγοριοποίησης - categorization) είτε η αυτοματοποιημένη δόμηση συλλογών εγγράφων με την εύρεση ομάδων με παρόμοια έγγραφα (μέθοδοι ομαδοποίησης - clustering methods). Υπάρχουν επίσης μέθοδοι για την αυτοματοποιημένη εξαγωγή χρήσιμων μοτίβων πληροφοριών από συλλογές εγγράφων κειμένου (εξαγωγή πληροφοριών - information extraction), μέθοδοι για την οπτικοποίηση της εξόρυξης κειμένου, όπως και μέθοδοι για την ανίχνευση θεμάτων και τη σύνοψη κειμένων.

3.7.2.1 Ταξινόμηση - Classification

Η ταξινόμηση κειμένων έχει ως στόχο να αντιστοιχίσει έγγραφα κειμένου σε προκαθορισμένες κλάσεις (Mitchell 1997). Για παράδειγμα, η αυτόματη επισήμανση κάθε εισερχόμενης είδησης με μια ετικέτα όπως ‘αθλητισμός’, ’πολιτική’, ‘τέχνη’ κλπ. Μια εργασία ταξινόμησης δεδομένων ξεκινά με την εκπαίδευση ενός συνόλου D = (d1,…, dn) εγγράφων που έχουν ήδη αντιστοιχιστεί σε μια κλάση L ∈ L (π.χ. αθλητισμός, πολιτική). Στη συνέχεια στόχος είναι να καθοριστεί ένα μοντέλο ταξινόμησης f : D → L f(d) = L που είναι σε θέση να αντιστοιχίσει ένα νέο έγγραφο d στη σωστή κλάση. Για τη μέτρηση της απόδοσης ενός μοντέλου ταξινόμησης, ένας τυχαίος αριθμός εγγράφων που φέρουν ετικέτα παραμερίζεται και δε χρησιμοποιείται κατά την εκπαίδευση του μοντέλου. Έπειτα μπορούμε να ταξινομήσουμε τα έγγραφα αυτού του δοκιμαστικού συνόλου χρησιμοποιώντας το μοντέλο ταξινόμησης και να συγκρίνουμε τις εκτιμώμενες ετικέτες με τις πραγματικές ετικέτες. Ο αριθμός των σωστά ταξινομημένων εγγράφων σε σχέση με τον συνολικό αριθμό των εγγράφων ονομάζεται ακρίβεια του μοντέλου και είναι ένα πρώτο μέτρο απόδοσης.
Ωστόσο, συμβαίνει συχνά η κλάση στόχος να καλύπτει μόνο ένα μικρό ποσοστό των εγγράφων της συλλογής. Τότε θα έχουμε υψηλή ακρίβεια αν προσδώσουμε σε κάθε έγγραφο την εναλλακτική του κλάση. Για να αποφύγουμε αυτό το φαινόμενο χρησιμοποιούνται διαφορετικά μέτρα για τον προσδιορισμό της επιτυχίας της ταξινόμησης. Η ακρίβεια ποσοτικοποιεί τα ανακτημένα έγγραφα που είναι σχετικά, δηλαδή αυτά που ανήκουν στην κατηγορία στόχο. Η ανάκληση υποδεικνύει το ποσοστό των σχετικών εγγράφων που ανακτήθηκαν.

$$\text{precision} = \frac{\#\{relevant \cap retrieved\}}{\#retrieved} \qquad \text{recall} = \frac{\#\{relevant \cap retrieved\}}{\#relevant}$$

Είναι εμφανές ότι υπάρχει μια αντιστάθμιση μεταξύ ακρίβειας και ανάκλησης. Οι περισσότεροι ταξιμονητές προσδιορίζουν εσωτερικά κάποιο «βαθμό συμμετοχής» στην κατηγορία - στόχο. Αν μόνο έγγραφα με υψηλό βαθμό σχετικότητας εκχωρούνται στην κατηγορία - στόχο, η ακρίβεια είναι υψηλή. Ωστόσο, υπάρχει ο κίνδυνος πολλά σχετικά έγγραφα να έχουν

παραβλεφθεί, πράγμα που αντιστοιχεί σε χαμηλή ανάκληση. Όταν από την άλλη η αναζήτηση είναι πιο εξονυχιστική η ανάκληση αυξάνεται όμως αυτό κοστίζει στην ακρίβεια του μοντέλου. Για τον σκοπό αυτό υπάρχει η βαθμολογία F η οποία λαμβάνει υπόψη τους παράγοντες της ακρίβειας και της ανάκλησης με σκοπό τη μέτρηση της συνολικής απόδοσης των ταξινομητών. (Hotho et al., 2005)

$$F = \frac{2}{1/recall + 1/precision}$$

3.7.2.2 Ομαδοποίηση - Clustering

Η μέθοδος ομαδοποίησης μπορεί να χρησιμοποιηθεί για την εύρεση ομάδων εγγράφων με παρόμοιο περιεχόμενο. Το αποτέλεσμα της ομαδοποίησης είναι συνήθως ένας διαχωρισμός, που ονομάζεται επίσης και ομαδοποίηση P, ένα σύνολο συμπλεγμάτων P. Κάθε σύμπλεγμα αποτελείται από έναν αριθμό εγγράφων d. Τα αντικείμενα, στην προκειμένη περίπτωση τα έγγραφα, ενός συμπλέγματος πρέπει να είναι παρόμοια μεταξύ τους και παράλληλα ανόμοια με τα έγγραφα άλλων συμπλεγμάτων. Τις περισσότερες φορές η ποιότητα των ομαδοποιήσεων θεωρείται ποιοτικότερη όταν τα περιεχόμενα των εγγράφων μέσα σε ένα σύμπλεγμα είναι όμοια ενώ μεταξύ των συστάδων παρατηρείται μια ανομοιογένεια. Οι αλγόριθμοι ομαδοποίησης υπολογίζουν τα συμπλέγματα με βάση τα χαρακτηριστικά των δεδομένων με τα οποία τροφοδοτούνται οι αλγόριθμοι και ανάλογα με τις μετρήσεις (αν)ομοιότητας που πραγματοποιούνται. Παρόλα αυτά, η ιδέα για το πως πρέπει να είναι η ιδανική ομαδοποίηση διαφέρει μεταξύ περιπτώσεων χρήσης ή ακόμα και μεταξύ της οπτικής των χρηστών. Κάποιος μπορεί να χειραγωγήσει τα αποτελέσματα ενός αλγορίθμου ομαδοποίησης χρησιμοποιώντας μόνο υποσύνολα χαρακτηριστικών ή προσαρμόζοντας τα χρησιμοποιούμενα μέτρα ομοιότητας. Ο βαθμός στον οποίο τα αποτελέσματα του αλγορίθμου συμβαδίζουν με την αντίληψη του χρήστη εκτιμάται με τη βοήθεια διάφορων μέτρων αξιολόγησης. Υπάρχουν δύο τρόποι αξιολόγησης των αποτελεσμάτων ομαδοποίησης. Από τη μια είναι οι στατιστικές μετρήσεις που μπορούν να χρησιμοποιηθούν για να περιγράψουν τις ιδιότητες των αποτελεσμάτων ομαδοποίησης, ενώ από την άλλη με βάση κάποια δεδομένη ταξινόμηση η οποία μπορεί να θεωρηθεί ως σημείο αναφοράς συγκρίνονται τα αποτελέσματα ομαδοποίησης με τη δεδομένη ταξινόμηση. (Hotho et al., 2005)

3.7.2.3 Εξαγωγή Πληροφοριών - Information Extraction

Αν και αναφερθήκαμε ήδη συνοπτικά στο κομμάτι της εξαγωγής πληροφοριών σε προηγούμενη υποενότητα θα δούμε μερικές ακόμα πληροφορίες, αναλύοντας και ένα παράδειγμα για να γίνει πιο κατανοητό.

Το κείμενο φυσικής γλώσσας περιέχει πολλές πληροφορίες που δεν είναι άμεσα προσβάσιμες για αυτόματη ανάλυση από υπολογιστή. Εντούτοις, οι υπολογιστές μπορούν να χρησιμοποιηθούν για να ‘κοσκινίσουν’ μεγάλο αριθμό κειμένων και να εξάγουν πληροφορίες από λέξεις, φράσεις ή αποσπάσματα. Ως εκ τούτου η εξαγωγή πληροφοριών μπορεί να θεωρηθεί σαν μια περιορισμένη μορφή πλήρους κατανόησης της φυσικής γλώσσας, όπου γνωρίζουμε εκ των προτέρων ποιες σημασιολογικές πληροφορίες αναζητούμε. Ο κύριος στόχος λοιπόν αυτής της διαδικασίας είναι η εξαγωγή τμημάτων του κειμένου με σκοπό να προσδώσουμε σε αυτά συγκεκριμένα χαρακτηριστικά. Για παράδειγμα, ποιες πληροφορίες μπορούν να εξαχθούν διαβάζοντας την είδηση: "Ο Robert L. James, πρόεδρος και διευθύνων

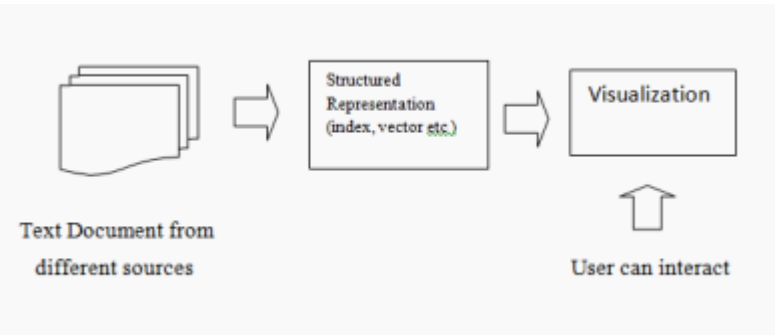
σύμβουλος της McCann Erickson, πρόκειται να συνταξιοδοτηθεί την 1η Ιουλίου. Θα αντικατασταθεί από τον John J. Donner, Jr., ο οποίος είναι επικεφαλής επιχειρησιακός διευθυντής των πρακτορείων". Σε αυτή την περίπτωση θα προσδιοριστούν οι παρακάτω πληροφορίες:

- οργανισμός (McCann-Erickson),
- θέση (διευθύνων σύμβουλος),
- ημερομηνία (1 Ιουλίου),
- όνομα εξερχόμενου ατόμου (Robert L. James),
- όνομα εισερχόμενου ατόμου (John J. Donner, Jr.)

Το έργο της εξαγωγής πληροφοριών διαχωρίζεται σε μια σειρά από στάδια επεξεργασίας που περιλαμβάνουν τη διακριτικοποίηση, τον διαχωρισμό προτάσεων, την ανάθεση μερών του λόγου και την ταυτοποίηση ονοματικών οντοτήτων όπως π.χ. ονόματα προσώπων, τοποθεσιών, οργανισμών κλπ. Σε υψηλότερο επίπεδο τώρα, φράσεις και προτάσεις πρέπει να αναλύονται και να ερμηνεύονται σημασιολογικά πριν ενσωματωθούν στη διαδικασία. Τελικά, οι απαραίτητες πληροφορίες όπως "θέση" και "όνομα εισερχόμενου ατόμου" εισάγονται στη βάση δεδομένων. Ακόμα και τα πιο ακριβή συστήματα εξαγωγής πληροφοριών συχνά απαιτούν τη χειροκίνητη παραμετροποίηση από τον χρήστη για την επεξεργασία της φυσικής γλώσσας, ωστόσο αξίζει να σημειωθεί ότι σε αρκετά από τα στάδια της διαδικασίας εξόρυξης δεδομένων έχει σημειωθεί σημαντική πρόοδος. (Hotho et al, 2005)

3.7.2.4 Οπτικοποίηση - Visualization

Η γραφική απεικόνιση των πληροφοριών θεωρείται συχνά πληρέστερη και πιο κατανοητή από ότι η πληροφόρηση που μπορούν να μας παρέχουν οι γραπτές περιγραφές, καθιστώντας την οπτικοποίηση ιδιαίτερα βοηθητική στην εξόρυξη μεγάλων συλλογών εγγράφων. Πολλές από τις προσεγγίσεις που αναπτύχθηκαν με σκοπό την εξόρυξη κειμένου είναι εμπνευσμένες από μεθόδους που είχαν προταθεί στους τομείς της διερευνητικής ανάλυσης δεδομένων (explorative data analysis), της οπτικοποίησης πληροφοριών (information visualization) και της εξόρυξης δεδομένων με οπτική απεικόνιση (visual data mining).
Στην εξόρυξη κειμένου ή στα συστήματα ανάκτησης πληροφοριών οι μέθοδοι οπτικοποίησης μπορούν να βελτιώσουν και να απλοποιήσουν την ανακάλυψη ή την εξαγωγή σχετικών προτύπων ή πληροφοριών. Πρόκληση σε αυτόν τον τομέα αποτελεί το γεγονός πως είναι εξαιρετικά δύσκολη η οπτικοποίηση αφηρημένων κειμενικών πληροφοριών ειδικά για τις συλλογές εγγράφων, όπως επίσης και οι ιδιαίτερες απαιτήσεις κατά τον σχεδιασμό της διεπαφής χρήστη (Hotho et al., 2005). Στην εικόνα που ακολουθεί φαίνονται τα βήματα κατά τη διαδικασία της οπτικοποίησης.



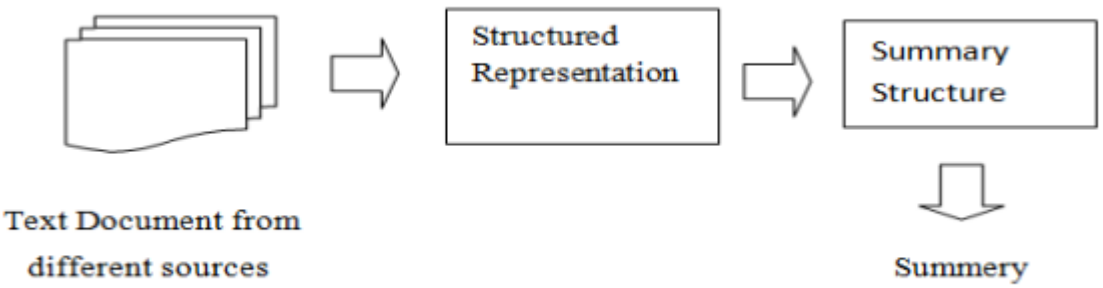
Εικόνα 6: Οπτικοποίηση

3.7.2.5 Σύνοψη Κειμένου - Text Summarization

Η σύνοψη κειμένου χρησιμοποιείται για να μειώσει το μήκος και τις λεπτομέρειες που περιέχει ένα έγγραφο διατηρώντας παράλληλα τα πιο σημαντικά σημεία και το γενικό του νόημα. Είναι χρήσιμη για τον χρήστη καθώς τον βοηθά να καταλάβει αν το κείμενο ανταποκρίνεται στις ανάγκες του εξοικονομώντας παράλληλα χρόνο. Παρόλο που οι υπολογιστές είναι σε θέση να αναγνωρίζουν τοποθεσίες, άτομα, και διάφορες οντότητες είναι δύσκολο να εκπαιδεύσεις κάποιον αλγόριθμο ώστε να πραγματοποιεί εννοιολογική ανάλυση και να ερμηνεύει το βαθύτερο νόημα ενός κειμένου. Αν και οι περισσότερες προσεγγίσεις εστιάζουν στην ιδέα της εξαγωγής μεμονωμένων καίριων προτάσεων, οι οποίες αργότερα θα αποτελέσουν και την περίληψη του κειμένου, υπάρχει ωστόσο και πιο εξελιγμένη προσέγγιση του προβλήματος με την εξαγωγή σημασιολογικών πληροφοριών από τα έγγραφα και χρήση αυτών μετέπειτα για τη δημιουργία μιας σύνοψης (Leskovec et al. 2004). Η διαδικασία της σύνοψης κειμένου αποτελείται από τα ακόλουθα βήματα (βλέπε εικόνα 7):

1. Κατά την προεπεξεργασία το πρωτότυπο κείμενο αποκτά μια δομημένη παρουσίαση
2. Για τη μετάβαση από την αρχική δομή στην περιληπτική δομή του κειμένου εφαρμόζονται κάποιοι αλγόριθμοι
3. Στο τελικό στάδιο δημιουργίας έχουμε τη συνοπτική παρουσίαση του κειμένου η οποία έχει προκύψει από την περιληπτική δομή του προηγούμενου βήματος.

(VijayGaikwad S., Chaugule A. & Patil P., 2014)



Εικόνα 7: Σύνοψη Κειμένου

3.8 Εφαρμογές Ανάλυσης Κειμένου

Όπως είδαμε στις προηγούμενες ενότητες η ανάλυση κειμένου μπορεί να βρει χρήση σε διάφορες επιστήμες και να εξυπηρετήσει διάφορους σκοπούς, ανάλογα με τις επιθυμίες και τους στόχους του εκάστοτε χρήστη. Είναι μια διαδικασία η οποία ξεκινάει με την

προεπεξεργασία της συλλογής κειμένων που επιθυμούμε να αναλύσουμε και ανάλογα με τον σκοπό της έρευνας - αναζήτησης που πρόκειται να πραγματοποιήσουμε μπορεί να καταλήξει σε ομαδοποίηση κειμένων, εύρεση σχετικών όρων, εξαγωγή πληροφοριών, εύρεση μοτίβων κ.α. Υπάρχουν πολλές εφαρμογές (λογισμικά, πλατφόρμες, ιστοσελίδες), εμπορικές και μη, που μπορούν να χρησιμοποιηθούν στην ανάλυση κειμένου. Σε αυτή την υποενότητα θα επιχειρήσουμε να κάνουμε μια κατηγοριοποίηση αυτών των εφαρμογών και να αναφέρουμε κάποια παραδείγματα για την κάθε κατηγορία. Έτσι λοιπόν έχουμε τον διαχωρισμό σε έξι (6) κατηγορίες:

3.8.1 Λογισμικά προετοιμασίας και καθαρισμού κειμένων

Τα κειμενικά δεδομένα δεν είναι πάντα στη μορφή που χρειάζεται ώστε να προχωρήσουμε σε περαιτέρω ανάλυση τους. Αρκετά συχνά κρίνεται αναγκαία η απομάκρυνση αχρείαστων τμημάτων όπως διάφορα σύμβολα ή σημεία στίξης, URLs, λέξεις που επαναλαμβάνονται ή στοπ λέξεις (λέξεις που δεν έχει νόημα να συμμετέχουν στην ανάλυση όπως άρθρα, σύνδεσμοι κ.λπ.) και κενά διαστήματα ή γραμμές. Εφαρμογές επεξεργασίας κειμένου όπως το Microsoft Word ή το Google Docs μας δίνουν τη δυνατότητα να πραγματοποιήσουμε τέτοιου είδους καθαρισμό στα κείμενα μας ωστόσο υπάρχουν και κάποιες πιο εξειδικευμένες εφαρμογές για αυτό τον σκοπό.

3.8.2 Λογισμικά ανάλυσης κειμένων γενικής χρήσης

Πρόκειται για λογισμικά τα οποία δεν επικεντρώνονται αποκλειστικά σε κάποιο βήμα της ανάλυσης κειμένου και η ανάλυση που πραγματοποιούν μπορεί να έχει ποιοτικά και ποσοτικά χαρακτηριστικά. Κάποια λογισμικά της κατηγορίας μάλιστα υποστηρίζουν όλα τα βήματα της διαδικασίας εξόρυξης δεδομένων όπως προετοιμασία δεδομένων, έλεγχος εγκυρότητας δεδομένων, ανάλυση και οπτικοποίηση δεδομένων.

3.8.3 Λογισμικά ποιοτικής ανάλυσης δεδομένων

Σε αυτή την κατηγορία έχουμε λογισμικά που εστιάζουν περισσότερο στην κατανόηση και τον εντοπισμό θεμάτων και μοτίβων μέσα σε ένα κείμενο ή μια συλλογή κειμένων. Αν και κάποια από αυτά διαθέτουν ενότητες για στατιστική ανάλυση ή τη δυνατότητα διασύνδεσης με στατιστικά εργαλεία ο κύριος ρόλος τους είναι η ποιοτική και όχι η ποσοτική ανάλυση των κειμένων. Η ποιοτική ανάλυση του περιεχομένου των κειμένων ξεπερνά την απλή καταμέτρηση ή τον εντοπισμό λέξεων και επικεντρώνεται περισσότερο στη νοηματική και θεματική ανάλυση του περιεχομένου των κειμένων. Η ανάλυση συναισθήματος (sentiment analysis), η εξαγωγή χαρακτηριστικών (feature extraction), η θεματική ανάλυση (thematic analysis), η ταξινόμηση θεμάτων (topic classification) και η ανάλυση μεταφορών (metaphor analysis) είναι κάποιες από τις δυνατότητες που παρέχουν αυτά τα λογισμικά.

3.8.4 Λογισμικά εξόρυξης γνώμης

Αν και αυτή η κατηγορία θα μπορούσε να έχει ενσωματωθεί στην προηγούμενη εντούτοις οι εφαρμογές που ανήκουν εδώ μπορούν να αποτελέσουν μια ξεχωριστή κατηγορία. Πρόκειται για εφαρμογές που εστιάζουν περισσότερο στο κομμάτι της εξόρυξης γνώμης (opinion mining) ή αλλιώς στην ανάλυση συναισθήματος (sentiment analysis). Ουσιαστικά πρόκειται για μια

τεχνική ανάλυσης κειμένων που χρησιμοποιεί υπολογιστική γλωσσολογία (computational linguistics) και επεξεργασία φυσικής γλώσσας (natural language processing) με σκοπό να αυτοματοποιήσει τον εντοπισμό και την εξαγωγή συναισθημάτων ή γνώμης από ένα κείμενο.

3.8.5 Λογισμικά ευρετηριοποίησης όρων και λέξεων κλειδιών

Αυτά τα λογισμικά παρέχουν τη δυνατότητα αναζήτησης όρων ή φράσεων, όπως επίσης και τη δυνατότητα δημιουργίας ευρετηρίου όρων. Είναι λογισμικά που δεν έχουν τη δυνατότητα σύνθετης ανάλυσης κειμένου και αρκούνται σε απλές λίστες συχνότητας εμφάνισης λέξεων, στην ευρετηριοποίηση όρων και στην εύρεση λέξεων ή φράσεων σχετικών με την αναζήτηση του χρήστη.

3.8.6 Λογισμικά οπτικοποίησης

Τα λογισμικά οπτικοποίησης χρησιμοποιούνται σε συνδυασμό με άλλα λογισμικά ανάλυσης κειμένου (ή μπορεί ακόμα και να αποτελούν ενότητα κάποιου λογισμικού ανάλυσης κειμένου) με σκοπό την οπτικοποίηση των αποτελεσμάτων της ανάλυσης. Αυτού του τύπου τα λογισμικά αναπτύσσονται ταχέως και εξελίσσονται όλο και περισσότερο ώστε να παρέχουν όσο το δυνατό καλύτερη αναπαράσταση των αποτελεσμάτων της ανάλυσης που πραγματοποιήθηκε. Όπως ειπώθηκε και σε προηγούμενη ενότητα αποτελεί ιδιαίτερη πρόκληση ο σχεδιασμός της διεπαφής χρήστη καθώς είναι πιθανό να αλλοιωθεί το νόημα και η δύναμη της ανάλυσης μέσα από μια σύνθετη και ανακριβή οπτικοποίηση.
(Gabe Ignatow, Rada Mihalcea 2018)

ΚΑΤΗΓΟΡΙΑ	ΟΝΟΜΑΣΙΑ	ΚΟΣΤΟΣ	ΤΥΠΟΣ	ΒΡΕΙΤΕ ΕΔΩ
Λογισμικά προετοιμασίας και καθαρισμού κειμένων	**TextCleanr**	**Δωρεάν**	**Ιστοσελίδα**	**TextCleanr - Text Cleaner Tool**
	TextSoap	**Επί Πληρωμή**	**Εφαρμογή για Mac**	**TextSoap - Automate Your Text Cleanup**
	UltraEdit	**Επί Πληρωμή**	**Εφαρμογή για Win/Mac/Linux**	**UltraEdit Text Editor + Coding Software**

Λογισμικά ανάλυσης κειμένων γενικής χρήσης	**Leximancer**	**Επί Πληρωμή**	**Εφαρμογή για Mac/Win**	**Leximancer**
	RapidMiner	**Δωρεάν**	**Πλατφόρμα ανοικτού κώδικα**	**RapidMiner | Amplify the Impact of Your People, Expertise & Data**
	WordStat	**Επί Πληρωμή**	**Εφαρμογή για Win/Mac**	**Text Analysis & Mining Software**
Λογισμικά ποιοτικής ανάλυσης δεδομένων	**Atlas.ti**	**Επί Πληρωμή**	**Εφαρμογή για Mac/Win - Διαδικτυακή Εφαρμογή**	**ATLAS.ti - The Qualitative Data Analysis & Research Software (atlasti.com)**
	NVivo	**Επί Πληρωμή**	**Εφαρμογή για Mac/Win**	**Best Qualitative Data Analysis Software for Researchers**
	CATMA	**Δωρεάν**	**Εφαρμογή για Win/Mac/Linux**	**CATMA**
Λογισμικά εξόρυξης γνώμης	**MonkeyLearn**	**Επί Πληρωμή**	**Διαδικτυακή Πλατφόρμα**	**MonkeyLearn - Text Analytics**
	SAS VISUAL TEXT ANALYTICS	**Επί Πληρωμή**	**Διαδικτυακή Πλατφόρμα**	**SAS: Analytics, Artificial Intelligence and Data Management | SAS**
Λογισμικά ευρετηριοποίησης όρων και λέξεων κλειδιών	**TextSTAT**	**Δωρεάν**	**Εφαρμογή για Windows, GNU/Linux και Mac**	**TextSTAT :: Niederländische Philologie FU Berlin (fu-berlin.de)**

	Wmatrix	**Δωρεάν**	**Διαδικτυακή Εφαρμογή**	**Wmatrix corpus analysis and comparison tool (lancs.ac.uk)**
Λογισμικά οπτικοποίησης	**TagCrowd**	**Δωρεάν**	**Ιστοσελίδα**	**TagCrowd: create your own word cloud from any text**
	*** Λογισμικά που έχουν συμπεριληφθεί σε άλλες κατηγορίες παρέχουν επίσης τη δυνατότητα οπτικοποίησης (RapidMiner, Atlas.ti, NVivo, SAS)**			

Πίνακας 1: Παραδείγματα Εφαρμογών Ανάλυσης Κειμένου

4.Εφαρμογή Ανάλυσης Κειμένου Texter

4.1 Εισαγωγή

Ο Texter είναι μια διαδικτυακή εφαρμογή ανάλυσης κειμένου που δημιουργήθηκε στα πλαίσια της παρούσας εργασίας. Έχει δημιουργηθεί αποκλειστικά με τη χρήση της γλώσσας προγραμματισμού PHP, ενώ σαν μέσο αποθήκευσης και ανάλυσης των κειμένων που εισάγονται στην εφαρμογή χρησιμοποιεί τη βάση δεδομένων PostgreSQL. Για τη συγγραφή του κώδικα χρησιμοποιήθηκε ο επεξεργαστής κώδικα Visual Studio Code και η εφαρμογή δεν είναι ανεβασμένη σε κάποιον εξυπηρετητή για κοινή χρήση, αλλά λειτουργεί τοπικά με τη χρήση του XAMPP. Επίσης, αξίζει να αναφερθεί η χρήση του εργαλείου Composer για την εγκατάσταση κάποιων βιβλιοθηκών της PHP που υποστηρίζουν λειτουργίες όπως την εξαγωγή κειμένου από .pdf αρχεία, την επιβεβαίωση νέου λογαριασμού από τους χρήστες και τη λειτουργία forgot password. Σε αυτή την ενότητα θα δούμε και θα αναλύσουμε τις δυνατότητες του Texter και εν συνεχεία θα αναφερθούμε στις δυσκολίες και τους περιορισμούς που προέκυψαν κατά τη δημιουργία της εφαρμογής.

4.2 Δυνατότητες του Texter

Για την καλύτερη κατανόηση των δυνατοτήτων της εφαρμογής θα χωρίσουμε την παρουσίαση τους σε δύο κατηγορίες:

4.2.1 Γενικά Χαρακτηριστικά

Σε αυτή την κατηγορία ανήκουν τα χαρακτηριστικά της εφαρμογής που δεν έχουν να κάνουν με την κύρια λειτουργικότητα της εφαρμογής η οποία είναι η ανάλυση κειμένου.

- Σύστημα Login: για να χρησιμοποιήσει την εφαρμογή κάποιος χρήστης πρέπει πρώτα να εισάγει τα διαπιστευτήρια του (όνομα χρήστη/κωδικό).
- Εγγραφή Χρήστη: σε περίπτωση νέου χρήστη θα πρέπει να δημιουργηθεί νέος λογαριασμός ο οποίος για να είναι λειτουργικός θα πρέπει να ταυτοποιηθεί μέσω ενός συνδέσμου που θα αποσταλεί στο email που δήλωσε ο χρήστης κατά την εγγραφή του.
- Forgot Password Feature: Αν ο χρήστης έχει λογαριασμό, όμως έχει ξεχάσει τον κωδικό πρόσβασης του, μπορεί να τον ανακτήσει χρησιμοποιώντας το email που είχε δηλώσει κατά την εγγραφή του.
- Αλλαγή Κωδικού Πρόσβασης: Δίνεται η δυνατότητα στον χρήστη να αλλάξει τον κωδικό πρόσβασης του στην εφαρμογή.

- Πληροφορίες Λογαριασμού: Ο χρήστης μπορεί να δει κάποιες βασικές πληροφορίες για τον λογαριασμό του όπως όνομα χρήστη, email, αριθμό κειμένων που έχει ανεβάσει στον λογαριασμό του, κατάσταση λογαριασμού.

4.2.2 Κύρια Χαρακτηριστικά

Εδώ θα αναφερθούμε στα κύρια χαρακτηριστικά της εφαρμογής που αφορούν την εισαγωγή, την επεξεργασία και την ανάλυση των κειμένων.

- Εισαγωγή Κειμένων:
  - ➢ Χειροκίνητη εισαγωγή κειμένου: Ο χρήστης μπορεί είτε να γράψει εξ αρχής κάποιο κείμενο και να το ανεβάσει στην εφαρμογή είτε να κάνει αντιγραφή - επικόλληση από κάποια πηγή.
  - ➢ Ανέβασμα αρχείων από τον υπολογιστή: Ο χρήστης μπορεί να επιλέξει κάποια αρχεία από τον υπολογιστή του και να τα ανεβάσει στην εφαρμογή. Κάθε ανέβασμα μπορεί να περιέχει ως και 50 αρχεία συνολικού μεγέθους 500 MB και οι επεκτάσεις αρχείων που γίνονται δεκτές είναι .txt, .docx και .pdf.

- Διαχείριση Κειμένων:

➢ Ο χρήστης μπορεί να δει τους τίτλους των κειμένων που έχει ανεβάσει στην εφαρμογή καθώς και κάποιες σύντομες πληροφορίες για αυτά, όπως σύνολο λέξεων και προτάσεων για το κάθε κείμενο και ημερομηνία καταχώρησης. Επίσης, υπάρχει ένας συνοπτικός πίνακας με πληροφορίες που αφορούν το σύνολο των κειμένων που ανήκουν στον χρήστη.

➢ Ο χρήστης έχει τη δυνατότητα να διαγράψει κάποιο κείμενο αν δεν επιθυμεί να το έχει πλέον στη συλλογή του.

➢ Ο χρήστης έχει τη δυνατότητα να δει και να τροποποιήσει τον τίτλο ή το περιεχόμενο κάποιου κειμένου.

- Ανάλυση Κειμένων:

Ο χρήστης καλείται να επιλέξει τα κείμενα που επιθυμεί να αναλύσει. Κατά τη διάρκεια αυτής της διαδικασίας ο χρήστης σχηματίζει μια λίστα με τα προς ανάλυση κείμενα στην οποία μπορεί να εισάγει ή να αφαιρέσει κείμενα ένα προς ένα είτε ομαδικά με τη χρήση φίλτρων. Υπάρχει η δυνατότητα φιλτραρίσματος λέξης (ο χρήστης πληκτρολογεί μία λέξη ή φράση και εμφανίζονται τα κείμενα που την περιέχουν) και ημερομηνίας (ο χρήστης εισάγει ένα εύρος ημερομηνιών και του εμφανίζονται τα κείμενα που έχουν καταχωρηθεί τη ζητούμενη περίοδο). Αφού σχηματίσει τη λίστα του ο χρήστης μπορεί να προχωρήσει στην ανάλυση όπου εμφανίζονται πλέον οι επιλεγμένοι τίτλοι και του δίνεται η δυνατότητα να κάνει ανάλυση για όλα τα επιλεγμένα κείμενα αλλά και για κάθε κείμενο μεμονωμένα.

➢ Ανάλυση σε όλα τα κείμενα:

- Ανάλυση Σχετικότητας: Ο χρήστης πληκτρολογεί μια λέξη ή φράση και για κάθε τίτλο εμφανίζεται ένα σκορ το οποίο υποδηλώνει πόσο σχετικό είναι το κείμενο με τη λέξη ή φράση που έχει αναζητήσει ο χρήστης.

- Λέξεις που εμφανίζονται σε όλα τα κείμενα προς ανάλυση: Εντοπίζονται οι κοινές λέξεις μεταξύ των κειμένων προς ανάλυση.

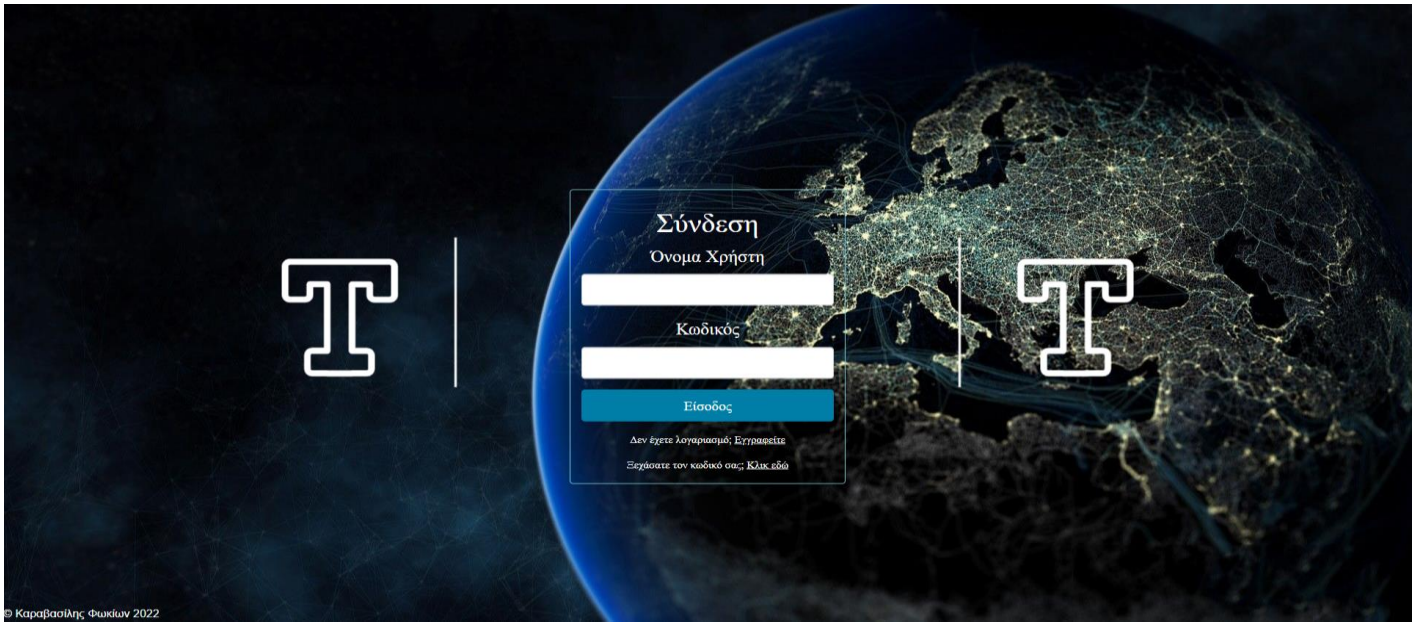
➢ Ανάλυση σε μεμονωμένο κείμενο:
- Ποσοστά: Εμφανίζει στον χρήστη κάποια ποσοστά σχετικά με τα stopwords, τις λέξεις που επαναλαμβάνονται, τις λέξεις που εμφανίζονται στο κείμενο μόνο μία φορά και τις μοναδικές λέξεις (ως μοναδικές λέξεις δεν ορίζονται οι λέξεις που εμφανίζονται μόνο μια φορά, περισσότερα στο εγχειρίδιο χρήσης). Επίσης, υπάρχει ένας πίνακας με την κατανομή των λέξεων ανάλογα με το μήκος τους σε χαρακτήρες.
- Stopwords: Εμφανίζει μια λίστα με τα stopwords που έχουν εντοπιστεί όπως επίσης το πλήθος τους και τον συνολικό αριθμό τους.
- Μοναδικές Λέξεις: Εμφανίζεται μια λίστα με τις ‘ρίζες’ των μοναδικών λέξεων. Ο χρήστης μπορεί να δει την πλήρη μορφή των λέξεων στις οποίες αναφέρεται η κάθε ‘ρίζα’, καθώς και να κάνει κλικ σε όποια λέξη επιθυμεί και να δει αποσπάσματα του κειμένου στα οποία εντοπίζεται η συγκεκριμένη λέξη. Υπάρχει η δυνατότητα φιλτραρίσματος ανάλογα με το μήκος της λέξης σε χαρακτήρες και τον αριθμό εμφανίσεων της λέξης.
- Κορυφαίες Λέξεις: Εμφανίζεται μια λίστα με τις ‘ρίζες’ των 10 λέξεων που εμφανίζονται πιο συχνά στο κείμενο. Ο χρήστης μπορεί να δει την πλήρη μορφή των λέξεων στις οποίες αναφέρεται η κάθε ‘ρίζα’. Υπάρχει η δυνατότητα φιλτραρίσματος ανάλογα με το μήκος της λέξης σε χαρακτήρες.

4.3 Εγχειρίδιο Χρήσης - Documentation

Στο εγχειρίδιο χρήσης του Texter θα δούμε αναλυτικά πως ο χρήστης μπορεί να δημιουργήσει ένα νέο λογαριασμό στην εφαρμογή, πως μπορεί να περιηγηθεί στην εφαρμογή και να κάνει χρήση των δυνατοτήτων της και θα αναλύσουμε τεχνικά χαρακτηριστικά, λειτουργίες και περιορισμούς, με σκοπό να γίνει πλήρως κατανοητό στο χρήστη πως προκύπτουν και τι σημαίνουν τα αποτελέσματα της ανάλυσης των κειμένων του.

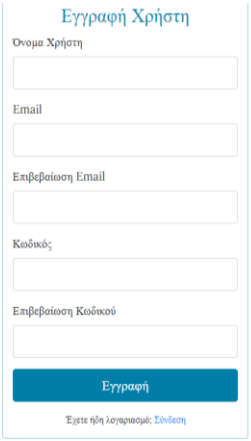
4.3.1 Δημιουργία νέου λογαριασμού χρήστη

Βήμα 1: Για να μπορέσει κάποιος να χρησιμοποιήσει την εφαρμογή κρίνεται απαραίτητο να δημιουργήσει πρώτα ένα προσωπικό λογαριασμό. Από την οθόνη σύνδεσης λοιπόν (βλ. εικόνα 8) πατήστε ‘Εγγραφείτε’.

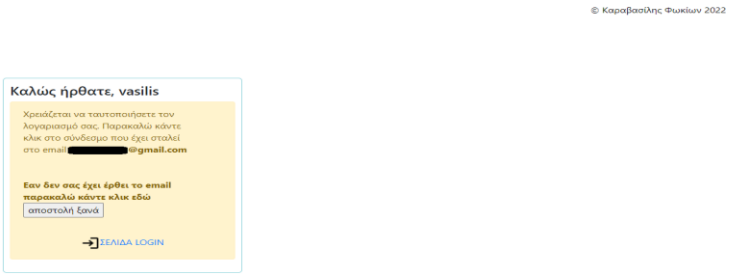


Εικόνα 8: Οθόνη Σύνδεσης Texter

Βήμα 2: Έχετε μεταφερθεί στην οθόνη συμπλήρωσης των προσωπικών στοιχείων του χρήστη (βλ. εικόνα 9). Εδώ καλείστε να συμπληρώσετε τα προσωπικά σας στοιχεία (email, όνομα χρήστη, κωδικό πρόσβασης). Σε περίπτωση που το όνομα χρήστη ή το email που θα εισάγετε χρησιμοποιείται ήδη, θα εμφανιστεί προειδοποιητικό μήνυμα στο άνω μέρος της φόρμας συμπλήρωσης των στοιχείων σας. Συμπληρώστε τη φόρμα και πατήστε ‘Εγγραφή’. Αν δεν υπάρξει κάποια προειδοποίηση ή κάποιο σφάλμα θα εμφανιστεί η οθόνη επιτυχούς δημιουργίας λογαριασμού (βλ. εικόνα 10).



Εικόνα 9: Οθόνη συμπλήρωσης προσωπικών στοιχείων



Εικόνα 10: Οθόνη επιτυχούς δημιουργίας λογαριασμού

Βήμα 3: Ο λογαριασμός σας έχει δημιουργηθεί επιτυχώς, όμως για να είναι λειτουργικός θα πρέπει να γίνει ταυτοποίηση του email που δηλώθηκε κατά την εγγραφή σας. Επισκεφτείτε τον λογαριασμό email σας και κάντε κλικ στο σύνδεσμο που σας έχει αποσταλεί. Σε περίπτωση που δε σας έχει αποσταλεί το σχετικό email, μπορείτε να εισάγετε τα διαπιστευτήρια σας στην οθόνη σύνδεσης (βλ. εικόνα 8) και θα σας εμφανιστεί και πάλι η οθόνη επιτυχούς δημιουργίας λογαριασμού (βλ. εικόνα 10). Από εκεί έχετε τη δυνατότητα να πατήσετε ‘αποστολή ξανά’. Όταν ταυτοποιήσετε τον λογαριασμό σας επιτυχώς θα εμφανιστεί η οθόνη επιτυχούς ταυτοποίησης (βλ. εικόνα 11). Κάνοντας κλικ στα πράσινα γράμματα όπως σας παραπέμπει και το γραπτό μήνυμα θα μεταφερθείτε εντός της εφαρμογής, είστε πλέον σε θέση να ξεκινήσετε να τη χρησιμοποιείτε!

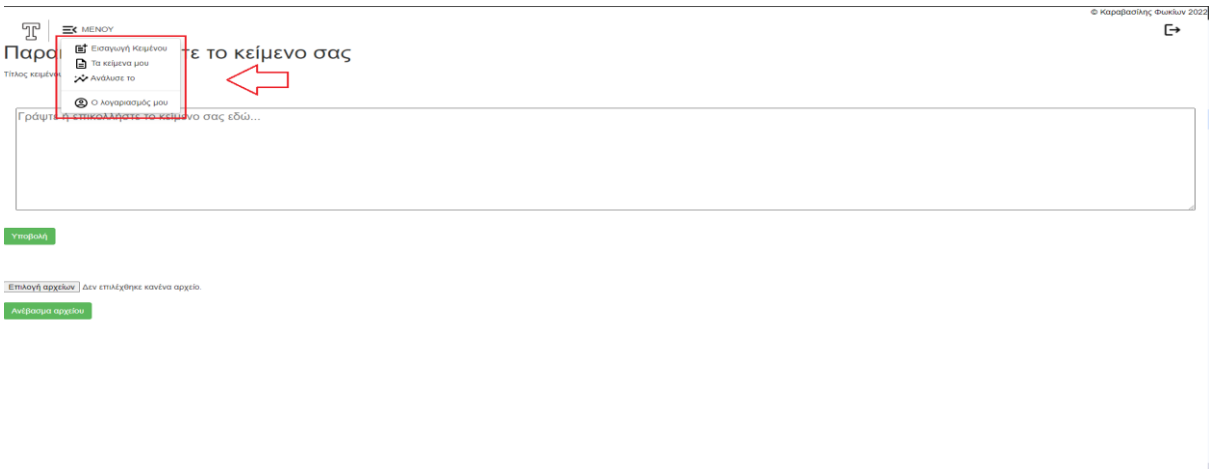
Εικόνα 11: Οθόνη επιτυχούς ταυτοποίησης

4.3.2 Περιήγηση εντός της εφαρμογής

Μπορείτε να περιηγηθείτε στην εφαρμογή από την μπάρα περιήγησης που βρίσκεται στο άνω μέρος της σελίδας της εφαρμογής (βλ. εικόνα 12).

Εικόνα 12: Μπάρα περιήγησης

Κάνοντας κλικ στο λογότυπο που βρίσκεται στο αριστερό μέρος της μπάρας περιήγησης θα μεταφέρεστε πάντα στην αρχική οθόνη της εφαρμογής η οποία είναι και η οθόνη της ενότητας ‘Εισαγωγή Κειμένου’ (βλ. εικόνα 14). Από το μενού (βλ. εικόνα 13) μπορείτε να εναλλάσσεστε μεταξύ των τεσσάρων ενοτήτων της εφαρμογής (θα τις δούμε αναλυτικότερα παρακάτω), ενώ στο δεξί μέρος της μπάρας βρίσκεται το εικονίδιο αποσύνδεσης το οποίο σας αποσυνδέει από τον λογαριασμό σας και σας μεταφέρει στην οθόνη σύνδεσης (βλ. εικόνα 8).



Εικόνα 13: ΜΕΝΟΥ

4.3.3 Παρουσίαση των τεσσάρων ενοτήτων της εφαρμογής

4.3.3.1 Ενότητα ‘Εισαγωγή Κειμένου’

Εικόνα 14:Οθόνη ενότητας ‘Εισαγωγή Κειμένου’

Όπως ειπώθηκε και παραπάνω αυτή η οθόνη αποτελεί παράλληλα και την αρχική οθόνη της εφαρμογής. Από εδώ μπορείτε να ανεβάζετε τα κείμενα σας είτε χειροκίνητα είτε μεταφορτώνοντας αρχεία κειμένου από τον υπολογιστή σας.

Χειροκίνητη Καταχώρηση

Για τη χειροκίνητη εισαγωγή κειμένου μπορείτε είτε να πληκτρολογήσετε εξ’ αρχής το κείμενο σας είτε να επικολλήσετε κάποιο απόσπασμα. Αφού συμπληρώσετε λοιπόν τίτλο και περιεχόμενο για το κείμενο σας πατήστε ‘Υποβολή’. Αν το κείμενο σας καταχωρηθεί επιτυχώς εμφανίζεται στο κάτω μέρος της σελίδας μήνυμα επιτυχούς καταχώρησης το οποίο θα παραμείνει για διάστημα δύο δευτερολέπτων.

Μεταφόρτωση Αρχείων

Πατήστε ‘Επιλογή αρχείων’ και επιλέξτε μαζικά τα αρχεία προς ανέβασμα, εν συνεχεία πατήστε ‘Ανέβασμα αρχείου’. Αν τα αρχεία σας καταχωρηθούν επιτυχώς θα εμφανιστεί μήνυμα επιτυχούς καταχώρησης στο κάτω μέρος της σελίδας. Προσοχή, δεν υπάρχει η δυνατότητα να προσθέσετε σε ένα ανέβασμα αρχεία που βρίσκονται αποθηκευμένα σε διαφορετικούς φακέλους στον υπολογιστή σας, συνίσταται λοιπόν να συγκεντρώσετε όλα τα αρχεία που επιθυμείτε να ανεβάσετε στην εφαρμογή σε έναν φάκελο ώστε να μπορέσετε να τα ανεβάσετε μαζικά.

Περιορισμοί και μηνύματα σφάλματος

- Η εφαρμογή δέχεται αρχεία με την επέκταση .docx, .txt και .pdf. Αν δοκιμάσετε να ανεβάσετε κάποιο αρχείο με διαφορετική επέκταση θα εμφανιστεί μήνυμα σφάλματος στο κάτω μέρος της σελίδας.
- Μπορείτε να ανεβάσετε μαζικά ως και 50 αρχεία συνολικού μεγέθους 500 MB. Αν ξεπεραστούν αυτά τα όρια εμφανίζεται μήνυμα σφάλματος στο κάτω μέρος της σελίδας.
- Δεν μπορείτε να καταχωρήσετε κείμενα με τον ίδιο τίτλο. Εμφανίζεται μήνυμα σφάλματος στο κάτω μέρος της σελίδας. Προσοχή διότι ο έλεγχος για την ύπαρξη κειμένων με τον ίδιο τίτλο είναι case - sensitive, που σημαίνει ότι ο τίτλος “Τίτλος 1”

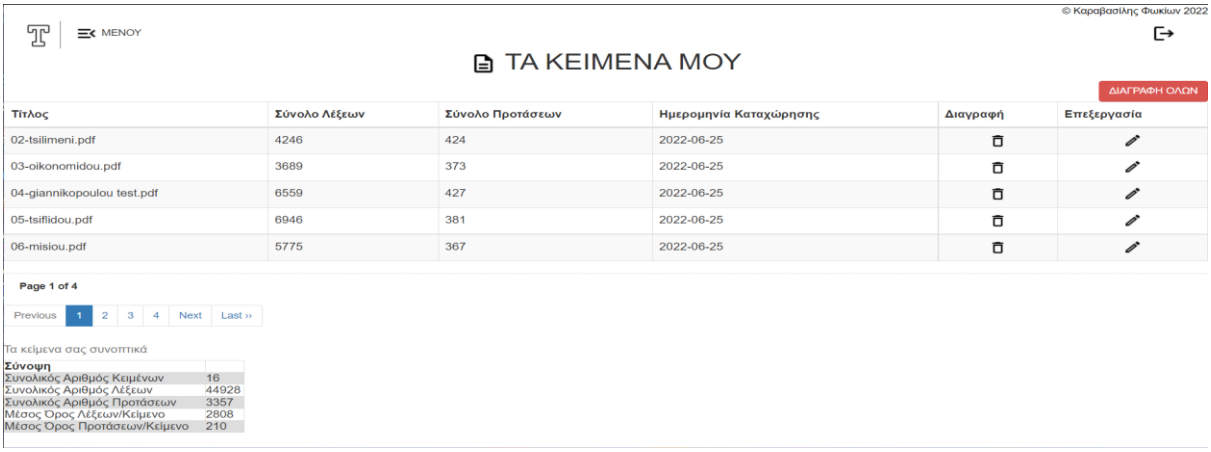
είναι διαφορετικός από τον τίτλο “τίτλος 1” καθώς στην πρώτη περίπτωση το γράμμα ‘τ’ είναι κεφαλαίο.
- Δεν μπορείτε να ανεβάσετε χειροκίνητα κάποιο κείμενο χωρίς να συμπληρώσετε τίτλο **και** περιεχόμενο. Εμφανίζεται μήνυμα σφάλματος στο κάτω μέρος της σελίδας.
- Σε περίπτωση που πατήσετε ‘Υποβολή’ ή ‘Ανέβασμα αρχείου’ έχοντας αφήσει κενά τον τίτλο και το περιεχόμενο στην πρώτη περίπτωση και δίχως να έχετε επιλέξει κάποιο αρχείο στη δεύτερη, θα εμφανιστεί μήνυμα σφάλματος στο κάτω μέρος της σελίδας.

Τι συμβαίνει κατά την καταχώρηση κειμένων στην εφαρμογή

Κατά την καταχώρηση των κειμένων σας γίνονται αυτοματοποιημένα οι απαραίτητες ενέργειες ώστε τα κείμενα να πάρουν την κατάλληλη μορφή πριν εισαχθούν στη βάση δεδομένων, όπου θα λάβουν χώρα κάποιες από τις μεθόδους προεπεξεργασίας κειμένων που αναφέραμε νωρίτερα σε άλλη ενότητα της εργασίας αυτής. Με τη χρήση λοιπόν regular expressions τροποποιείται η μορφή του κειμένου εξαλείφοντας παραπανίσια κενά και κενές γραμμές ενώ παράλληλα τα τρία θαυμαστικά και οι τρεις τελείες αντικαθίστανται με ένα/μία (αυτό συμβαίνει διότι ο αριθμός των προτάσεων προκύπτει από τη μέτρηση τους). Ουσιαστικά το κείμενο πρέπει να έρθει σε μία μορφή αλφαριθμητικού που ακολουθεί το μοτίβο λέξη - κενό - λέξη. Τότε και μόνο θα γίνει η εισαγωγή του κειμένου στη βάση δεδομένων η οποία αναλαμβάνει να διαχωρίσει το κείμενο σε λεξικογραφικές μονάδες (tokens), να αποκόψει τις καταλήξεις ή τα προθέματα των λεξικογραφικών μονάδων με τη χρήση σχετικού αλγορίθμου (snowball stemmer), να ευρετηριοποιήσει τους όρους του κειμένου και να ελέγξει το κείμενο για την ύπαρξη stopwords.

4.3.3.2 Ενότητα ‘Τα Κείμενα μου’

Σε αυτή την ενότητα μπορείτε να δείτε τους τίτλους των κειμένων που έχετε καταχωρήσει στην εφαρμογή, να τους επεξεργαστείτε ή να τους διαγράψετε. Τα κείμενα εμφανίζονται σε διάταξη σελιδοποίησης με την κάθε σελίδα να περιέχει πέντε (5) τίτλους. Δίπλα από κάθε τίτλο εμφανίζεται το σύνολο των λέξεων που περιέχει το εκάστοτε κείμενο, το σύνολο των προτάσεων, η ημερομηνία καταχώρησης, το πλήκτρο διαγραφής και τέλος το πλήκτρο επεξεργασίας του κειμένου. Στο κάτω αριστερά μέρος της σελίδας θα βρείτε ένα πινακάκι με συνοπτικές πληροφορίες για το σύνολο των κειμένων σας όπως ‘Συνολικός Αριθμός Κειμένων’, ‘Συνολικός Αριθμός Λέξεων’, ‘Συνολικός Αριθμός Προτάσεων’, ‘Μέσος Όρος Λέξεων/Κείμενο’ και ‘Μέσος Όρος Προτάσεων/Κείμενο’ (βλ. εικόνα 15).



Εικόνα 15: Οθόνη ενότητας 'Τα Κείμενα μου'

<u>Διαγραφή Κειμένου</u>

Για να διαγράψετε κάποιο κείμενο πατήστε το πλήκτρο με τον κουβά που εμφανίζεται στην τέταρτη στήλη δίπλα από τον τίτλο που επιθυμείτε να διαγράψετε. Μπορείτε επίσης να διαγράψετε μαζικά όλα τα κείμενα σας κάνοντας κλικ στο κόκκινο κουμπί 'ΔΙΑΓΡΑΦΗ ΟΛΩΝ' που βρίσκεται πάνω δεξιά από τον πίνακα που περιέχει τους τίτλους των κειμένων σας. Και στις δύο περιπτώσεις θα εμφανιστεί στο κάτω μέρος της οθόνης σας μήνυμα επιτυχούς διαγραφής.

<u>Επεξεργασία Κειμένου</u>

Για να τροποποιήσετε το περιεχόμενο ή τον τίτλο σε κάποιο από τα κείμενα σας πατήστε στο εικονίδιο με το μολύβι που εμφανίζεται στην πέμπτη στήλη δίπλα από τον τίτλο που επιθυμείτε να επεξεργαστείτε. Θα μεταφερθείτε στην οθόνη επεξεργασίας κειμένου (βλ. εικόνα 16). Πραγματοποιήστε τις αλλαγές σας στον τίτλο ή στο κυρίως κείμενο και πιέστε 'Υποβολή Αλλαγής'. Θα εμφανιστεί στο άνω μέρος της σελίδας μήνυμα επιτυχούς εφαρμογής των αλλαγών σας που θα σας υποδεικνύει και το που έγιναν οι αλλαγές αυτές, στον τίτλο, στο περιεχόμενο ή και στα δύο.

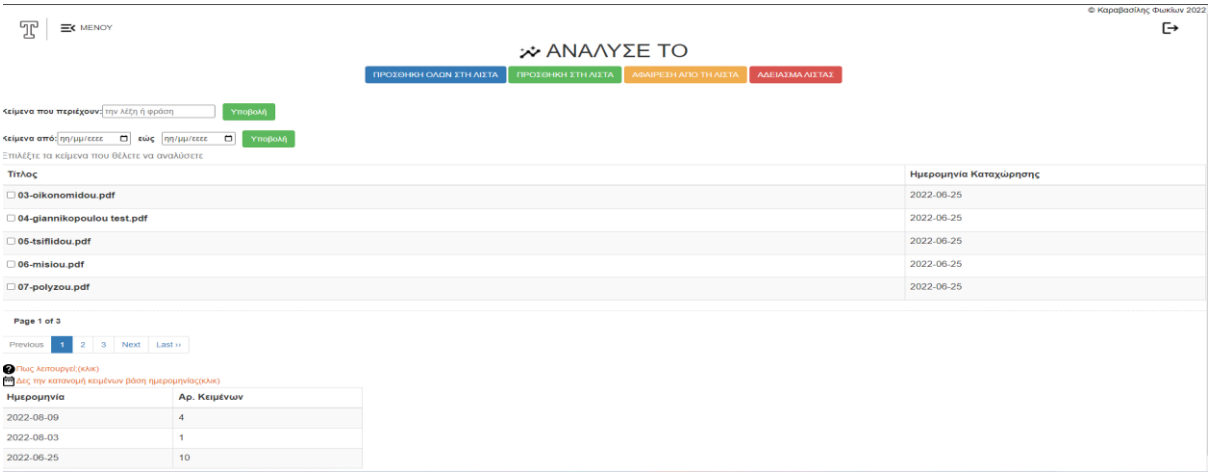
Εικόνα 16: Οθόνη επεξεργασίας κειμένου

<u>Περιορισμοί και μηνύματα σφάλματος</u>

- Αν δεν πραγματοποιήσετε κάποια αλλαγή και πατήσετε το πλήκτρο ‘Υποβολή Αλλαγής’ θα εμφανιστεί σχετικό μήνυμα σφάλματος στο άνω μέρος της σελίδας.
- Ενώ είναι επιτρεπτό να ανεβάσετε κάποιο αρχείο στην εφαρμογή που δεν έχει περιεχόμενο (π.χ. ένα κενό έγγραφο κειμένου), κατά την επεξεργασία του κειμένου δεν μπορείτε να πατήσετε ‘Υποβολή Αλλαγής’ ενώ ο τίτλος ή το κυρίως κείμενο είναι κενά καθώς θα εμφανιστεί μήνυμα σφάλματος στο άνω μέρος της σελίδας.
- Το σύνολο των προτάσεων προκύπτει από τη μέτρηση των τελείων και των θαυμαστικών. Καθώς αυτά δε σηματοδοτούν πάντα το τέλος μιας πρότασης και ίσως να έχουν διαφορετικό ρόλο μέσα σε ένα κείμενο είναι εμφανές πως ο αριθμός των προτάσεων που απαρτίζουν ένα κείμενο δεν είναι πάντα ακριβής.

4.3.3.3 Ενότητα ‘Ανάλυσέ το’

Πρόκειται για την πιο σημαντική ενότητα της εφαρμογής καθώς από εδώ πραγματοποιείται η επιλογή των κειμένων προς ανάλυση και έπειτα η ανάλυση τους. Έχετε τη δυνατότητα να φιλτράρετε τις επιλογές σας με τη χρήση του φίλτρου ‘ημερομηνίας εισαγωγής’ και του φίλτρου ‘λέξης ή φράσης’. Στο κάτω μέρος της οθόνης σας μπορείτε να πατήσετε στις φράσεις με τα πορτοκαλί γράμματα για να δείτε μια σύντομη περιγραφή σχετικά με το πως να πλαισιώσετε τη λίστα με τα κείμενα σας όπως επίσης και να δείτε την κατανομή των κειμένων σας ανάλογα με την ημερομηνία εισαγωγής τους στην εφαρμογή (βλ. εικόνα 17).



Εικόνα 17: Οθόνη ενότητας ‘Ανάλυσέ το’

Φίλτρο ‘λέξης ή φράσης’

Στο άνω μέρος της οθόνης σας όταν βρίσκεστε στην ενότητα ‘Ανάλυσέ το’ (βλ. εικόνα 17) βρίσκεται το φίλτρο λέξης ή φράσης. Πληκτρολογώντας μια λέξη ή μια ακολουθία λέξεων (το

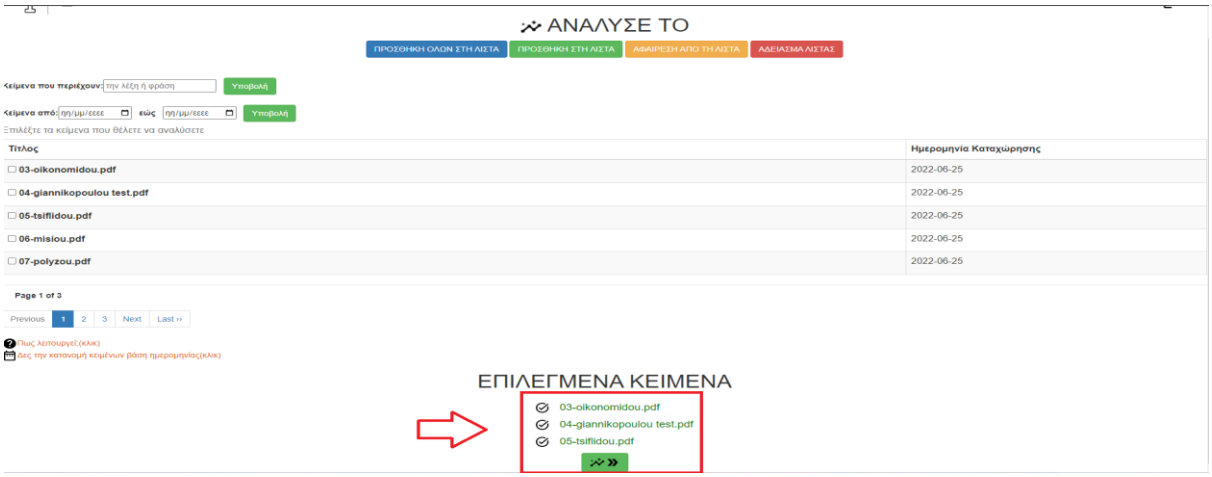
αναφέρω έτσι διότι δε χρειάζεται να έχει νοηματική συνέχεια) και πατώντας 'Υποβολή' θα σας εμφανιστούν τα κείμενα που είναι σχετικά με την αναζήτηση σας. Σε περίπτωση που πραγματοποιήσετε αναζήτηση για μεμονωμένη λέξη τα κείμενα που θα εμφανιστούν δε σημαίνει πως περιέχουν την ακριβή μορφή της λέξης που πληκτρολογήσατε καθώς το σύστημα νορμαλοποιεί τη λέξη. Σε περίπτωση που πραγματοποιήσετε αναζήτηση για μια ακολουθία λέξεων (ακολουθήστε το μοτίβο λέξη - κενό - λέξη) πάλι το σύστημα θα νορμαλοποιήσει την κάθε λέξη και θα σας εμφανίσει κείμενα που σχετίζονται με τις λέξεις αυτές, προσοχή δεν είναι απαραίτητο ότι τα φιλτραρισμένα κείμενα θα περιέχουν όλες τις λέξεις που πληκτρολογήσατε στη γραμμή αναζήτησης.

Φίλτρο ημερομηνίας

Ακριβώς κάτω από το φίλτρο 'λέξεις ή φράσης' βρίσκεται το φίλτρο ημερομηνίας. Συμπληρώνοντας ένα εύρος ημερομηνιών και πατώντας 'Υποβολή' θα σας εμφανιστούν τα κείμενα που έχουν εισαχθεί στην εφαρμογή στο ζητούμενο εύρος. Συμπληρώνοντας την ίδια ημερομηνία και στα δύο πεδία μπορείτε να κάνετε αναζήτηση για συγκεκριμένη ημέρα.

Δημιουργία της λίστας των κειμένων προς ανάλυση

Για να προχωρήσετε στην ανάλυση θα πρέπει πρώτα επιλέξετε τα κείμενα ή το κείμενο που επιθυμείτε να αναλύσετε. Επιλέγοντας τα κείμενα σας (θα δούμε αναλυτικά πως στη συνέχεια) οι τίτλοι τους θα εμφανίζονται στο κάτω μέρος της οθόνης σας (βλ. εικόνα 18).



Εικόνα 18: Λίστα των κειμένων προς ανάλυση

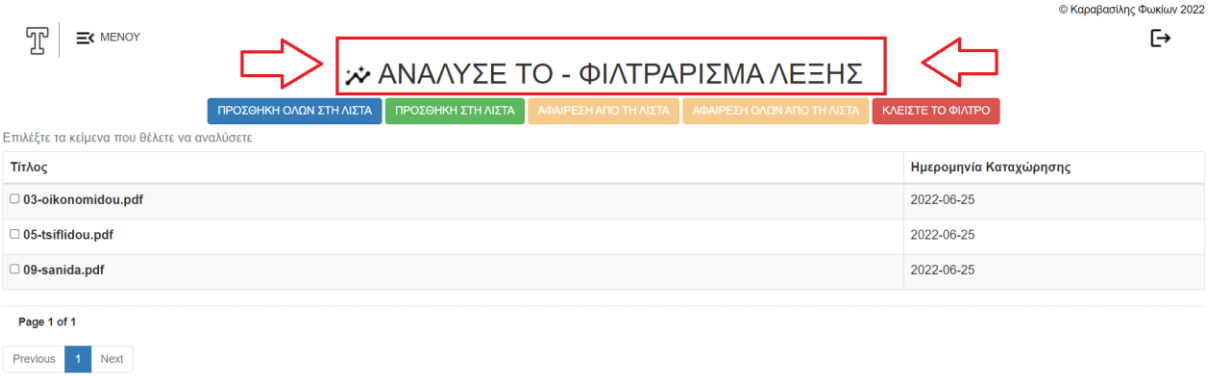
Για την πλαισίωση της λίστας σας χρησιμοποιείτε τα κουμπιά στο άνω μέρος της οθόνης της ενότητας 'Ανάλυσέ το' (βλ. εικόνα 17 ή 18). Ας δούμε όμως αναλυτικά τι κάνει το κάθε κουμπί:

- ΠΡΟΣΘΗΚΗ ΟΛΩΝ ΣΤΗ ΛΙΣΤΑ: προσθέτει το σύνολο των κειμένων σας στη λίστα των κειμένων προς ανάλυση.
- ΠΡΟΣΘΗΚΗ ΣΤΗ ΛΙΣΤΑ: αφού πρώτα τσεκάρετε έναν ή περισσότερους τίτλους πατώντας το θα προστεθούν στη λίστα σας. Προσοχή καθώς αλλάζοντας σελίδα (τα κείμενα εμφανίζονται με διάταξη σελιδοποίησης, πέντε κείμενα/σελίδα) οι τσεκαρισμένες επιλογές σας θα χαθούν αν δεν τις προσθέσετε στη λίστα σας.

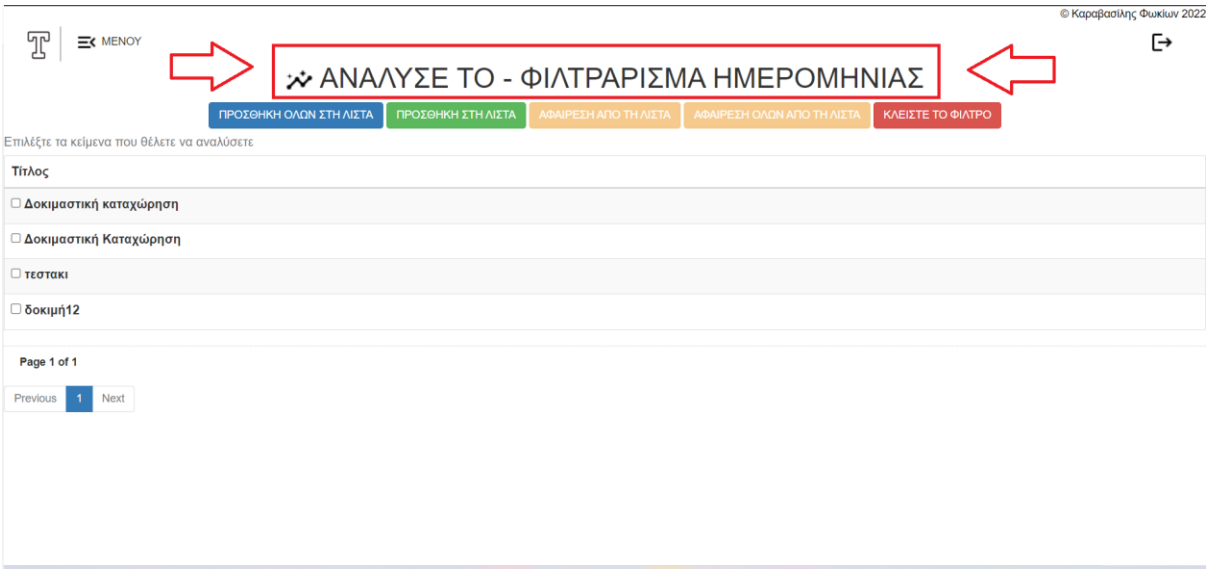
- ΑΦΑΙΡΕΣΗ ΑΠΟ ΤΗ ΛΙΣΤΑ: αφού πρώτα τσεκάρετε έναν ή περισσότερους τίτλους πατώντας το θα αφαιρεθούν από τη λίστα σας. Προσοχή καθώς αλλάζοντας σελίδα (τα κείμενα εμφανίζονται με διάταξη σελιδοποίησης, πέντε κείμενα/σελίδα) οι τσεκαρισμένες επιλογές σας θα χαθούν αν δεν τις αφαιρέσετε από τη λίστα σας.
- ΑΔΕΙΑΣΜΑ ΛΙΣΤΑΣ: αφαιρεί όλες τις επιλογές που έχετε κάνει και αδειάζει εντελώς τη λίστα σας.

Φυσικά, υπάρχει η δυνατότητα χρήσης των φίλτρων που αναφέραμε παραπάνω. Αφού συμπληρώσετε τα πεδία και πατήσετε ‘Υποβολή’ θα μεταφερθείτε στην οθόνη με τα αποτελέσματα της αναζήτησης σας. Καθώς η διεπαφή χρήστη μπορεί να σας μπερδέψει, συστήνεται να κοιτάτε στο άνω μέρος της σελίδας, όπου μπορείτε να δείτε αν βρίσκεστε στην επιλογή κειμένων ή στα αποτελέσματα του φίλτρου αναζήτησης (βλ. εικόνες 19 και 20). Η πλαισίωση της λίστα σας πραγματοποιείται με παρόμοιο τρόπο όπως και χωρίς τη χρήση φίλτρων. Εδώ πάλι έχουμε στο άνω μέρος της σελίδας φιλτραρισμένων αποτελεσμάτων (βλ. εικόνες 19 και 20) κάποια κουμπιά που σας βοηθούν να διαχειριστείτε τη λίστα σας. Πάμε να δούμε αναλυτικά τη χρήση τους:

- ΠΡΟΣΘΗΚΗ ΟΛΩΝ ΣΤΗ ΛΙΣΤΑ: προσθέτει το σύνολο των φιλτραρισμένων αποτελεσμάτων στη λίστα σας.
- ΠΡΟΣΘΗΚΗ ΣΤΗ ΛΙΣΤΑ: προσθέτει το επιλεγμένο φιλτραρισμένο κείμενο ή κείμενα στη λίστα σας. Προσοχή καθώς αλλάζοντας σελίδα (τα κείμενα εμφανίζονται με διάταξη σελιδοποίησης, πέντε κείμενα/σελίδα) οι τσεκαρισμένες επιλογές σας θα χαθούν αν δεν τις προσθέσετε στη λίστα σας.
- ΑΦΑΙΡΕΣΗ ΑΠΟ ΤΗ ΛΙΣΤΑ: αφαιρεί το επιλεγμένο φιλτραρισμένο κείμενο ή κείμενα από τη λίστα σας. Προσοχή καθώς αλλάζοντας σελίδα (τα κείμενα εμφανίζονται με διάταξη σελιδοποίησης, πέντε κείμενα/σελίδα) οι τσεκαρισμένες επιλογές σας θα χαθούν αν δεν τις αφαιρέσετε από τη λίστα σας.
- ΑΦΑΙΡΕΣΗ ΟΛΩΝ ΑΠΟ ΤΗ ΛΙΣΤΑ ΣΑΣ: αφαιρεί το σύνολο των φιλτραρισμένων αποτελεσμάτων από τη λίστα σας.
- ΚΛΕΙΣΤΕ ΤΟ ΦΙΛΤΡΟ: επιστρέφετε στην οθόνη επιλογής των κειμένων προς ανάλυση (δηλ. στην οθόνη της ενότητας ‘Ανάλυσέ το’).



Εικόνα 19: Οθόνη φίλτρου λέξης ή φράσης

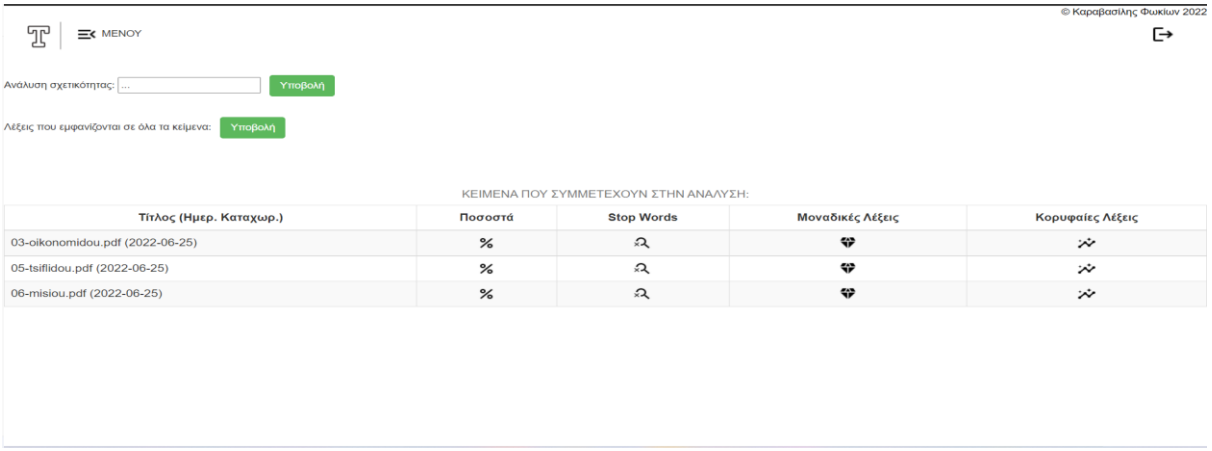


Εικόνα 20: Οθόνη φίλτρου ημερομηνίας

Καθώς δεν είναι ορατό στα στιγμιότυπα οθόνης παραπάνω, αξίζει να αναφερθεί πως αν έχετε προσθέσει κείμενα στη λίστα σας, αυτά είναι ορατά στο κάτω μέρος της οθόνης σας όταν χρησιμοποιείτε κάποιο από τα φίλτρα. Η ιδέα των φίλτρων μπορεί να σας μπερδεύει όμως μπορεί να φανεί ιδιαίτερα βοηθητική για κάποιον χρήστη ο οποίος έχει μεγάλο αριθμό κειμένων στη συλλογή του. Επιπρόσθετα μπορεί να απογειώσει τη διαλογή - επιλογή των κειμένων προς ανάλυση καθώς με τη χρήση των φίλτρων μπορούν να απαντηθούν ερωτήματα του τύπου: ποια κείμενα της συλλογής σας περιέχουν τη λέξη 1 αλλά δεν περιέχουν τη λέξη 2; ή ποια κείμενα περιέχουν τη λέξη 1 αλλά και τη λέξη 2; και άλλα σύνθετα ερωτήματα που αφορούν λέξεις, φράσεις και ημερομηνίες εισαγωγής.

Ανάλυση Κειμένων

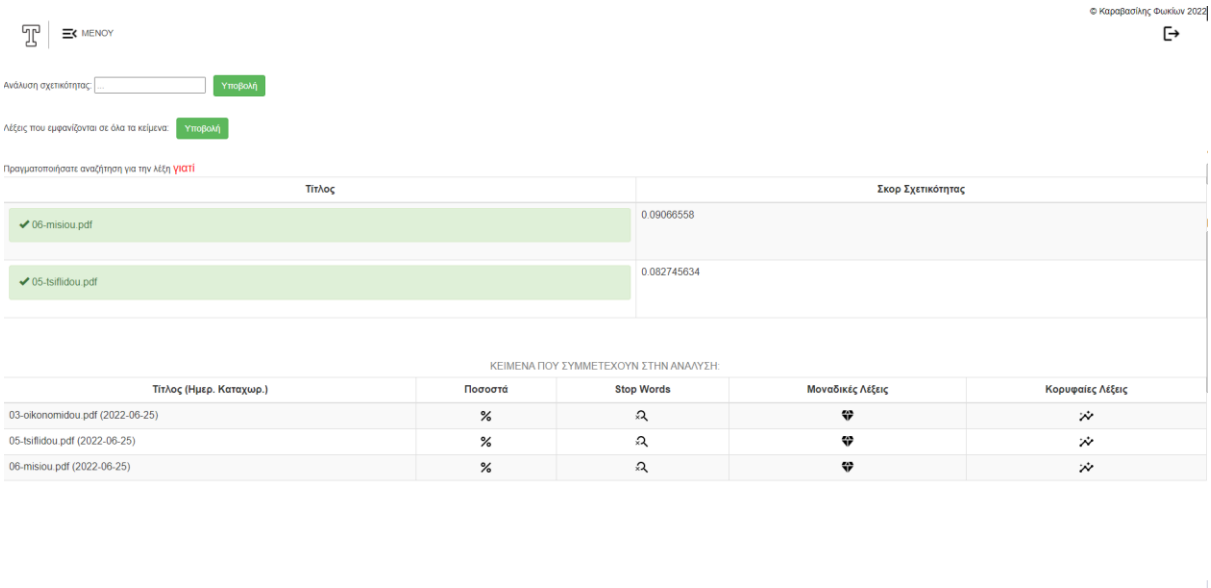
Αφού πλαισιώσετε τη λίστα σας με τα κείμενα ή το κείμενο που επιθυμείτε να αναλύσετε πατήστε το πράσινο κουμπί που βρίσκεται στο κάτω μέρος της οθόνης σας, κάτω από τη λίστα με τα επιλεγμένα κείμενα σας (βλ. εικόνα 18). Θα μεταφερθείτε στην οθόνη της ανάλυσης (βλ εικόνα 21) όπου σε περίπτωση που έχετε επιλέξει περισσότερα από ένα κείμενα σας δίνετε η δυνατότητα να δείτε κάποια στοιχεία για το σύνολο των επιλογών σας όπως επίσης και για κάθε κείμενο μεμονωμένα.



Εικόνα 21: Οθόνη ανάλυσης

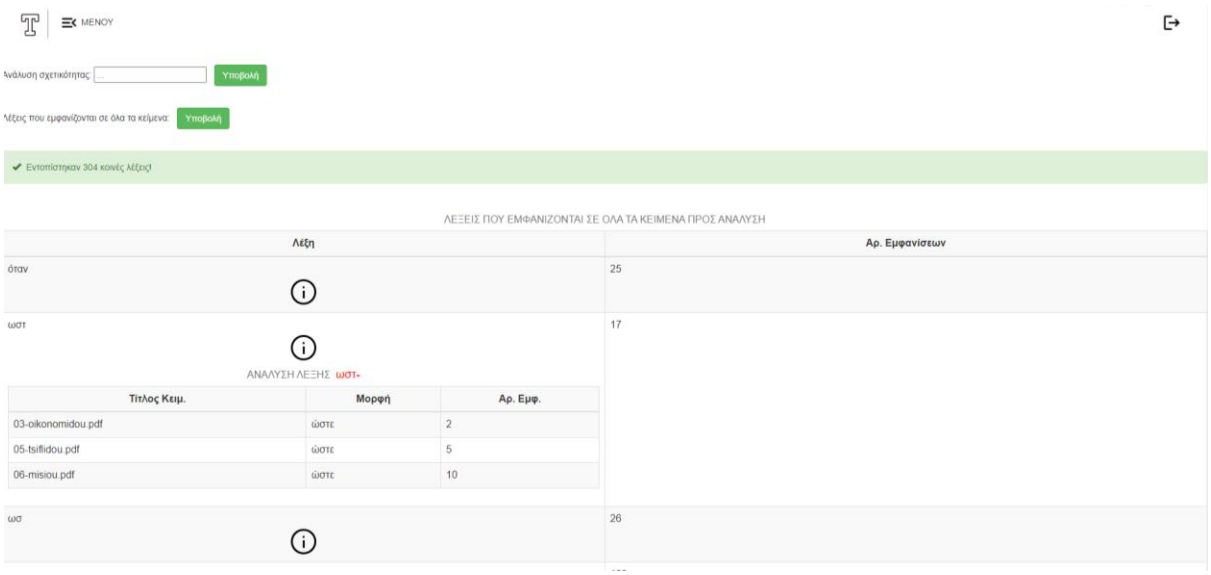
➢ **Ανάλυση σε όλα τα κείμενα**

- Ανάλυση σχετικότητας: στο άνω μέρος της οθόνης ανάλυσης (βλ. εικόνα 21) βρίσκεται το πεδίο της ανάλυσης σχετικότητας. Πληκτρολογώντας μια λέξη ή φράση (ακολουθώντας πάλι το μοτίβο λέξη - κενό - λέξη) θα σας εμφανίσει τους τίτλους των κειμένων που σχετίζονται με την αναζήτηση σας σε φθίνουσα κατάταξη με τον πιο σχετικό να βρίσκεται στην κορυφή της λίστας (βλ. εικόνα 22). Αν κανένα από τα κείμενα σας δεν είναι σχετικό με την αναζήτηση που πραγματοποιήσατε θα εμφανιστεί σχετικό μήνυμα. Η βαθμολογία σχετικότητας αποδίδεται από τη βάση δεδομένων και διαμορφώνεται ανάλογα με την έκταση του κειμένου, την απόσταση μεταξύ των ζητούμενων όρων μέσα στο κείμενο και τη συχνότητα εμφάνισης της λέξης - στόχου. (https://www.postgresql.org/docs/current/textsearch-controls.html)



Εικόνα 22: Ανάλυση σχετικότητας

- Λέξεις που περιέχονται σε όλα τα κείμενα: αν έχετε εισάγει στη λίστα σας περισσότερα από ένα κείμενα μπορείτε να πατήσετε υποβολή και να σας εμφανιστούν, αν υπάρχουν, οι λέξεις που περιέχονται σε όλα τα επιλεγμένα κείμενα σας. Πάνω ακριβώς από τον πίνακα που περιέχει τις λέξεις που εμφανίζονται σε όλα τα κείμενα προς ανάλυση θα εμφανιστεί ο συνολικός αριθμός αυτών των λέξεων ενώ δίπλα από κάθε λέξη υπάρχει ένα εικονίδιο το οποίο και αν πατήσετε θα σας εμφανιστεί αναλυτικά η μορφή της λέξης (στον πίνακα εμφανίζονται τα λεξήματα, δηλ. το απόκομμα της λέξης, και όχι η πλήρης μορφή της κάθε λέξης) καθώς επίσης και ο ακριβής αριθμός εμφάνισης της σε κάθε κείμενο που έχετε επιλέξει προς ανάλυση (βλ. εικόνα 23).



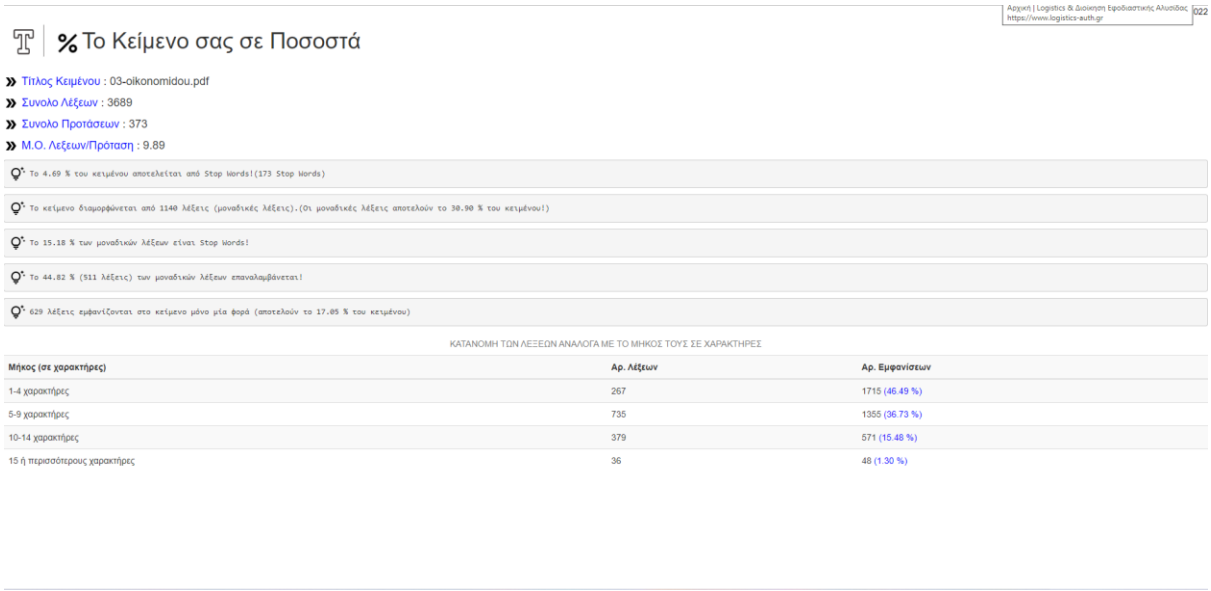
Εικόνα 23: Λέξεις που εμφανίζονται σε όλα τα κείμενα προς ανάλυση

➢ Ανάλυση σε μεμονωμένο κείμενο

***** Στο πάνω μέρος των σελίδων που ακολουθούν εμφανίζονται κάποιες γενικές πληροφορίες όπως ο τίτλος του κειμένου που αναλύεται, το σύνολο των λέξεων και των προτάσεων που περιέχει και ο μέσος όρος λέξεων/πρόταση.

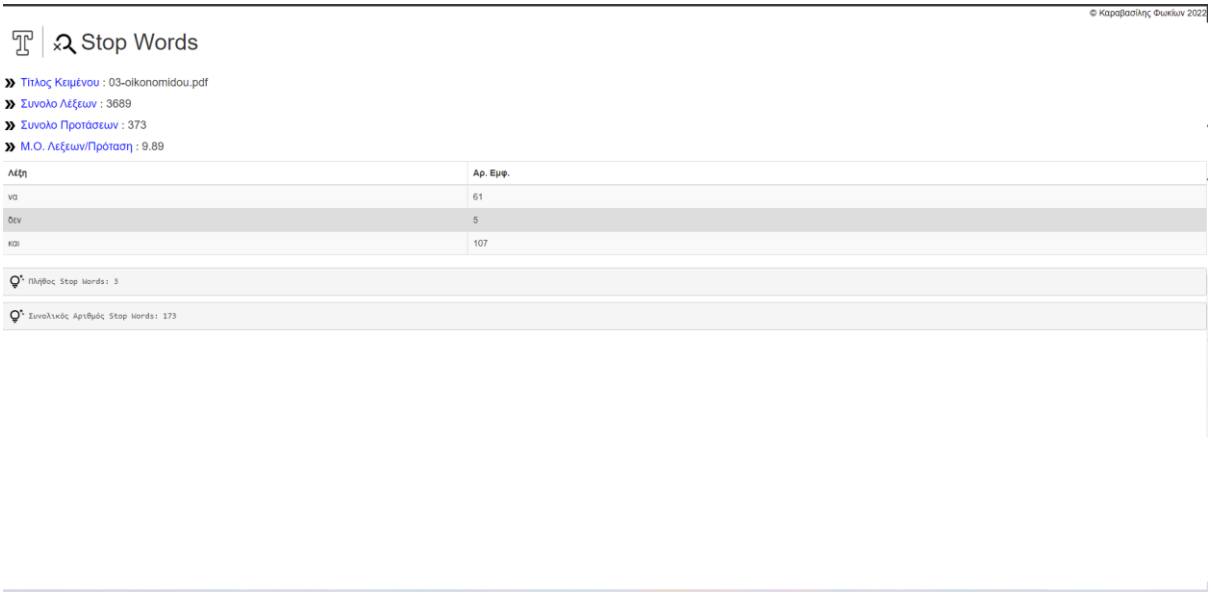
- Ποσοστά (βλ. εικόνα 24): εμφανίζονται κάποια ποσοστά που αφορούν:
  A. το ποσοστό του κειμένου που αποτελείται από stop words
  B. τι ποσοστό του κειμένου αποτελούν οι μοναδικές λέξεις (ως μοναδικές λέξεις ορίζονται τα λεξήματα που προκύπτουν μετά την εφαρμογή του αλγόριθμου αποκοπής καταλήξεων και όχι οι λέξεις που εμφανίζονται στο κείμενο μόνο μια φορά)
  C. τι ποσοστό των μοναδικών λέξεων είναι stop words
  D. τι ποσοστό των μοναδικών λέξεων επαναλαμβάνεται
  E. τι ποσοστό των συνολικών λέξεων του κειμένου εμφανίζονται στο κείμενο μόνο μια φορά

ενώ στο κάτω μέρος της οθόνης σας υπάρχει ένας πίνακας με την κατανομή των λέξεων ανάλογα με τους χαρακτήρες που τις αποτελούν.



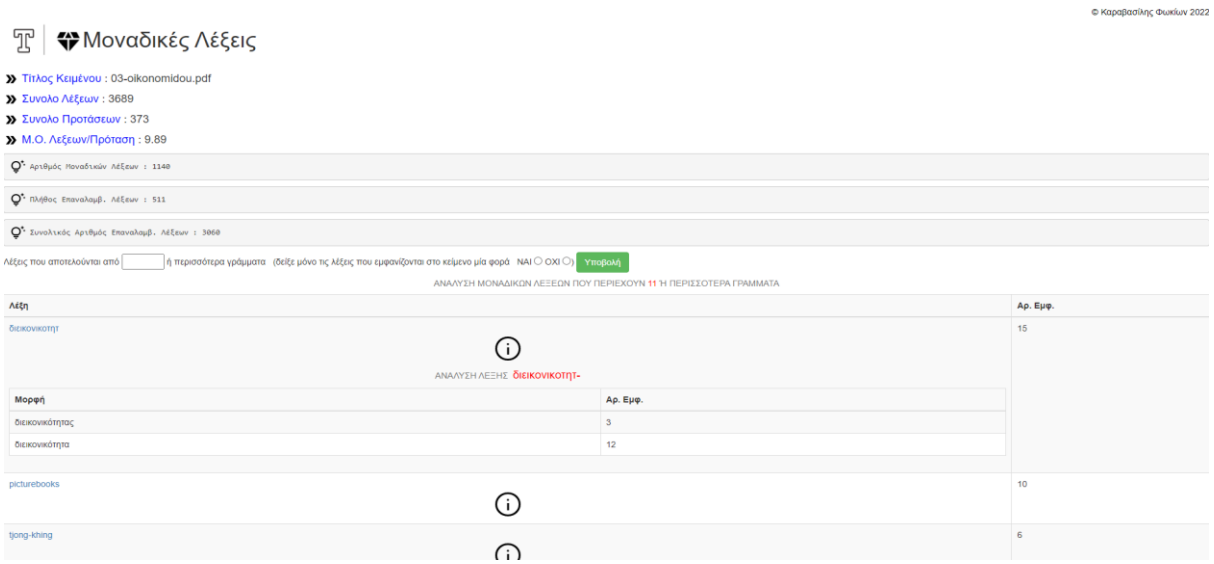
Εικόνα 24: Το κείμενο σας σε ποσοστά

- Stop Words (βλ. εικόνα 25): εμφανίζεται ένας πίνακας με τις στοπ λέξεις και τον αριθμό εμφανίσεων τους, όπως επίσης δίνεται το πλήθος τους και ο συνολικός αριθμός τους.

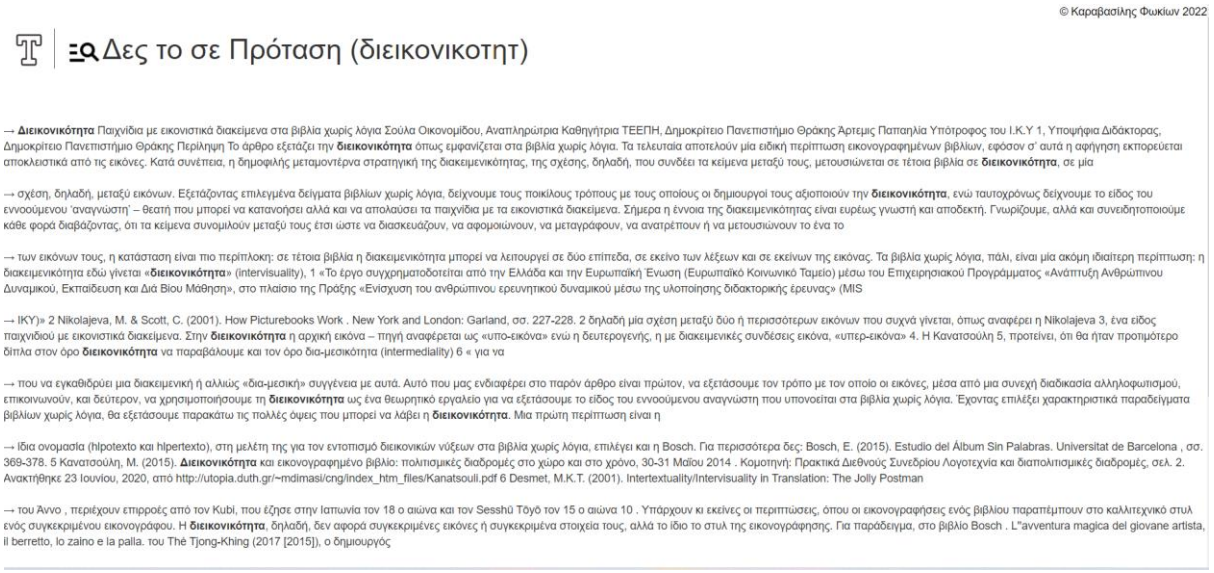


Εικόνα 25: Stop Words

- Μοναδικές Λέξεις (βλ. εικόνα 26): εμφανίζεται ο συνολικός αριθμός και το πλήθος των μοναδικών λέξεων (ορίσαμε τις μοναδικές λέξεις παραπάνω) όπως και ο συνολικός αριθμός των λέξεων που επαναλαμβάνονται. Μπορείτε να επιλέξετε τον αριθμό των χαρακτήρων και το αν εμφανίζονται στο κείμενο μόνο μια φορά οι μοναδικές λέξεις που αναζητάτε. Προσοχή διότι ο αριθμός των γραμμάτων που θα συμπληρώσετε αφορά τα λεξήματα και όχι την πλήρη μορφή της λέξης. Αφού συμπληρώσετε τα πεδία αυτά πατήστε ‘Υποβολή’. Θα εμφανιστεί ένας πίνακας με τα αποτελέσματα αναζήτησης που στη δεξιά στήλη περιέχει τα λεξήματα και στην αριστερή των αριθμό εμφάνισης τους. Δίπλα από κάθε λέξημα υπάρχει ένα εικονίδιο το οποίο και αν πατήσετε σας εμφανίζεται η πλήρης μορφή του λεξήματος. Επίσης, σας δίνεται η δυνατότητα να πατήσετε πάνω σε οποιοδήποτε λέξημα και να δείτε απόσπασμα από το κείμενο το οποίο περιέχει το συγκεκριμένο λέξημα (βλ. εικόνα 27).

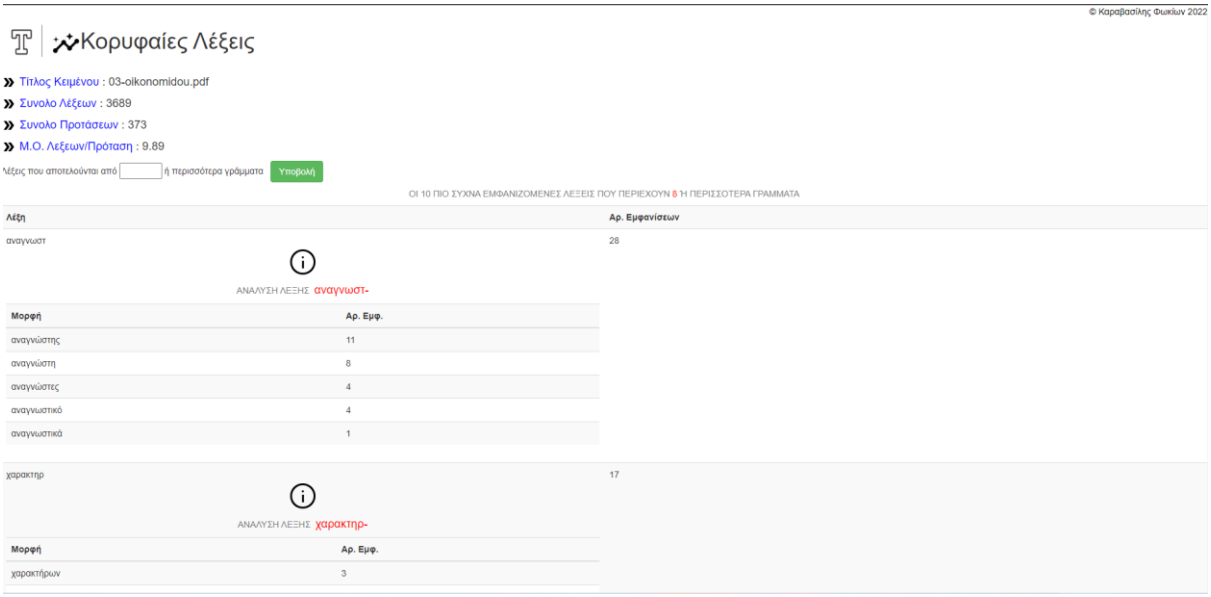


Εικόνα 26: Μοναδικές Λέξεις



Εικόνα 27: Δες το σε Πρόταση

- Κορυφαίες Λέξεις (βλ. εικόνα 28): Πρόκειται για τις δέκα λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης. Συμπληρώνοντας τον αριθμό των γραμμάτων που επιθυμείτε να έχουν οι κορυφαίες λέξεις που αναζητάτε και πατώντας 'Υποβολή' θα σας εμφανιστεί ένας πίνακας αποτελεσμάτων. Στη δεξιά στήλη εμφανίζονται τα λεξήματα ενώ στην αριστερή ο αριθμός εμφανίσεων τους. Και εδώ υπάρχει δίπλα από κάθε λέξημα ένα εικονίδιο στο οποίο και αν πατήσετε θα σας εμφανιστεί η πλήρης μορφή του λεξήματος.



Εικόνα 28: Κορυφαίες Λέξεις

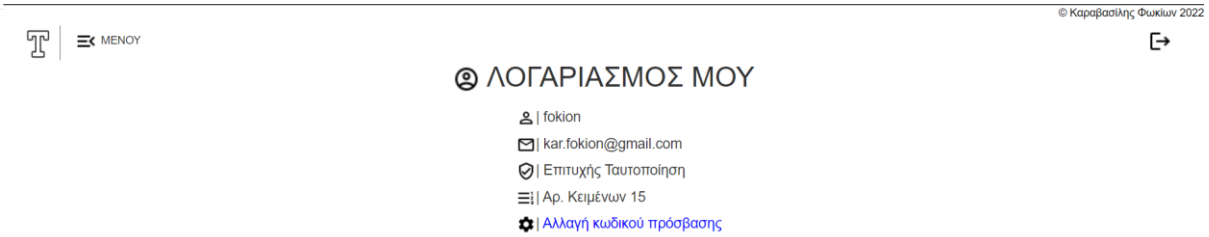
Περιορισμοί και μηνύματα σφάλματος

- Κατά τη διαδικασία πλαισίωσης της λίστας σας με τα κείμενα προς ανάλυση μπορούν να προκύψουν διάφορα μηνύματα σφάλματος που όμως όλα είναι κατατοπιστικά σχετικά με τις ενέργειες στις οποίες πρέπει να προβείτε για να σταματήσετε να τα λαμβάνετε.
- Αν κάποιο από τα πεδία που χρησιμοποιούνται για να φιλτράρουν τα αποτελέσματα αναζήτησης σας δε συμπληρωθεί και πατήσετε υποβολή θα εμφανιστεί μήνυμα σφάλματος.
- Στις μοναδικές λέξεις είναι απαραίτητη η συμπλήρωση και των δυο πεδίων για να σας εμφανιστούν αποτελέσματα. Αν επιθυμείτε να δείτε όλες τις μοναδικές λέξεις συμπληρώστε στα πεδία τον αριθμό ‘1’ και ‘ΟΧΙ’.
- Οποιαδήποτε αναζήτηση φέρει ως απάντηση μεγάλο αριθμό αποτελεσμάτων θα αργήσει να εμφανίσει το σύνολο των αποτελεσμάτων και ίσως ο browser που χρησιμοποιείτε να φορτώσει για αρκετά δευτερόλεπτα πριν εμφανίσει το πλήρες αποτέλεσμα. Για παράδειγμα, αν ζητήσετε να δείτε όλες τις μοναδικές λέξεις ενός μεγάλου σε έκταση κειμένου τότε θα χρειαστεί να περιμένετε μερικά δευτερόλεπτα, ανάλογα με την έκταση του κειμένου.

4.3.3.4 Ενότητα ‘Ο λογαριασμός μου’

Σε αυτή την ενότητα μπορείτε να δείτε γενικές πληροφορίες του λογαριασμού σας όπως όνομα χρήστη, email, αν είναι επιτυχώς ταυτοποιημένος και τον αριθμό των κειμένων που έχετε

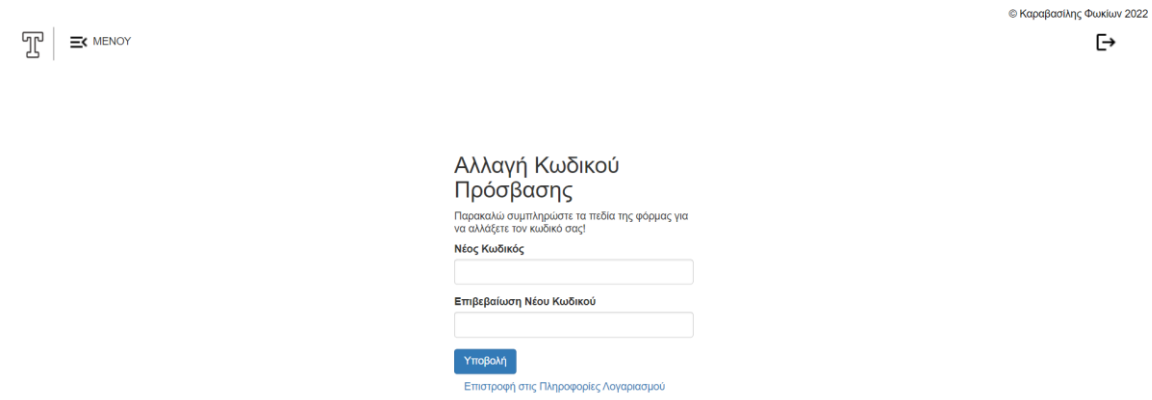
εισάγει στην εφαρμογή. Μπορείτε επίσης να αλλάξετε τον κωδικό του λογαριασμού σας (βλ. εικόνα 29).



Εικόνα 29: Οθόνη ενότητας 'Ο λογαριασμός μου'

Αλλαγή Κωδικού Πρόσβασης

Κάνοντας κλικ στην επιλογή 'Αλλαγή κωδικού πρόσβασης' θα μεταφερθείτε στη σελίδα αλλαγής κωδικού πρόσβασης (βλ. εικόνα 30). Συμπληρώστε και στα δύο πεδία τον νέο σας κωδικό πρόσβασης και πατήστε 'Υποβολή'. Θα εμφανιστεί μήνυμα επιτυχούς αλλαγής στο πάνω αριστερά μέρος της οθόνης σας και θα αποσυνδεθείτε από τον λογαριασμό σας. Πλέον για να συνδεθείτε στην εφαρμογή θα πρέπει να βάλετε τον νέο κωδικό πρόσβασης σας.



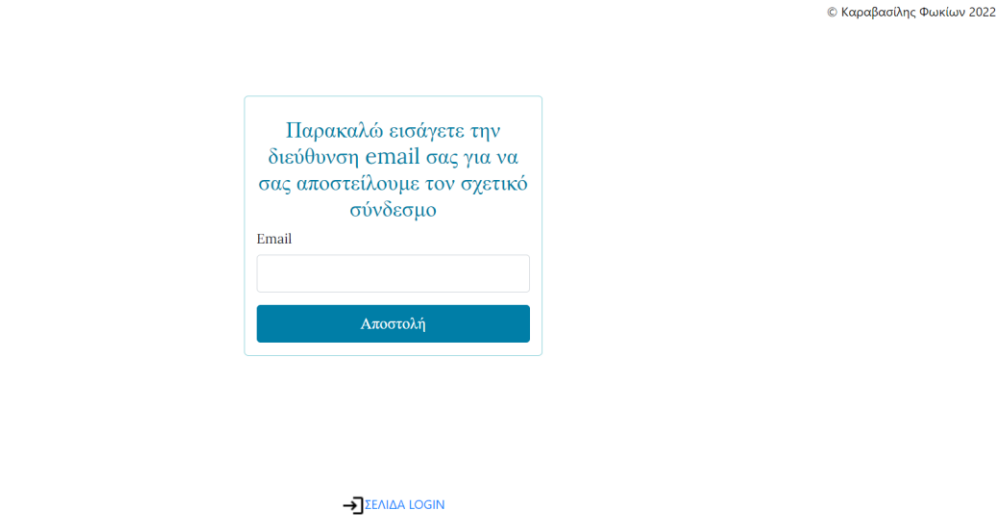
Εικόνα 30: Αλλαγή Κωδικού Πρόσβασης

Περιορισμοί και μηνύματα σφάλματος

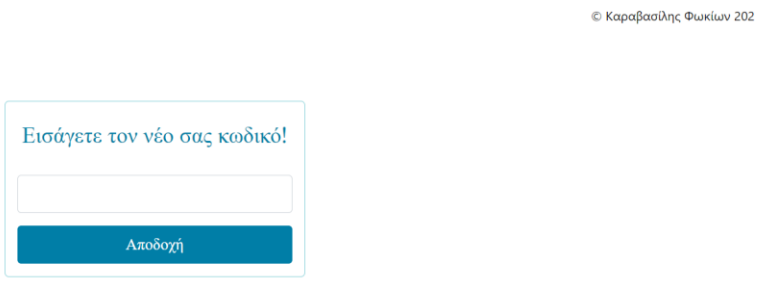
- Αν οι κωδικοί που θα συμπληρώσετε στα δύο πεδία δεν ταυτίζονται θα λάβετε το ανάλογο μήνυμα σφάλματος
- Δεν υπάρχουν περιορισμοί σχετικά με τη δομή ή το μέγεθος του κωδικού πρόσβασης σας.
- Ο νέος κωδικός μπορεί να είναι ίδιος με τον παλιό, δε θα υπάρξει κάποιο ενημερωτικό - προειδοποιητικό μήνυμα.

4.3.4 Forgot Password Feature

Σε περίπτωση που έχετε ξεχάσει τον κωδικό πρόσβασης σας μπορείτε:
Βήμα 1: να πατήσετε στην οθόνη εισόδου δίπλα από το ‘Ξεχάσατε τον κωδικό σας;’ (βλ. εικόνα 8).
Βήμα 2:Θα σας ζητηθεί να εισάγετε το email που είχατε δηλώσει κατά την εγγραφή σας ώστε να σας αποσταλεί εκεί ένας σύνδεσμος για την ανάκτηση του κωδικού σας (βλ. εικόνα 31). Μόλις πατήσετε ‘Αποστολή’ θα σας εμφανιστεί μήνυμα επιτυχίας. Σε περίπτωση που δε λάβετε το σχετικό email παρακαλώ επαναλάβετε τη διαδικασία.
Βήμα 3:Πατήστε στο σύνδεσμο που σας έχει αποσταλεί και εισάγετε τον κωδικό πρόσβασης που επιθυμείτε να έχετε από εδώ και στο εξής. Στη συνέχεια πατήστε ‘Αποδοχή’ (βλ. εικόνα 32). Ο κωδικός σας έχει ανακτηθεί επιτυχώς αν φορτώσει ο browser και στο πάνω αριστερά μέρος της σελίδας που θα φορτώσει έχει εμφανιστεί μήνυμα επιτυχίας καθώς και ένας σύνδεσμος που σας παραπέμπει να τον πατήσετε για να συνδεθείτε στην εφαρμογή.



Εικόνα 31: Forgot Password Feature



Εικόνα 32: Forgot Password Feature 2

Περιορισμοί και μηνύματα σφάλματος

- Ο κωδικός σας μπορεί να είναι ίδιος με αυτόν που έχετε ξεχάσει, δε θα προκύψει κάποιο σφάλμα.
- Δεν υπάρχουν περιορισμοί - παροτρύνσεις για τη δομή και το μέγεθος του κωδικού σας.
- Θα εισάγετε τον κωδικό σας μόνο μία φορά (δεν υπάρχει πεδίο επιβεβαίωσης κωδικού) φροντίστε να έχετε πληκτρολογήσει σωστά τον κωδικό σας.

4.4 Δυσκολίες και Περιορισμοί

Όπως είναι φυσικό, κατά τη διάρκεια δημιουργίας της εφαρμογής υπήρξαν κάποιες δυσκολίες και αργότερα με το πέρας της συγγραφής του κώδικα και την ολοκλήρωση της εφαρμογής παρατηρήθηκαν κάποια σημεία στα οποία υστερεί. Σε αυτό το κομμάτι της διπλωματικής θα δούμε εν συντομία τις δυσκολίες που αντιμετωπίστηκαν και τους περιορισμούς που έχει η εφαρμογή (επιφυλάσσομαι για πιθανούς περιορισμούς οι οποίοι δεν έχουν εντοπιστεί ακόμη).

- Η πρώτη μου επαφή με τον προγραμματισμό ήταν στο παρόν μεταπτυχιακό πρόγραμμα, επομένως οι προκλήσεις ήταν αρκετές όταν ανέλαβα να κάνω μια διπλωματική εργασία που συνεπαγόταν τη δημιουργία μιας διαδικτυακής εφαρμογής.
- Η βάση δεδομένων που χρησιμοποιούσε η εφαρμογή στα πρώτα στάδια της δημιουργίας της ήταν η MariaDB. Μετά από προτροπή του επιβλέπων καθηγητή μου

και αφού εξετάσαμε από κοινού τις δυνατότητες που προσφέρει η PostgreSQL στο τομέα της ευρετηριοποίησης κειμένων και της ανάκτησης και διαχείρισης πληροφοριών από κειμενικά δεδομένα, αποφασίσαμε να αλλάξουμε τη βάση δεδομένων με την οποία θα συνεργάζεται η εφαρμογή.

- Κάποιες αλλαγές στις ρυθμίσεις ασφαλείας της Google δημιούργησαν πρόβλημα στην ταυτοποίηση χρήστη και γενικά σε όποια λειτουργία της εφαρμογής συνεπάγεται την αποστολή email από την εφαρμογή σε πιθανό χρήστη.
- Η εφαρμογή είναι σχεδιασμένη να αναλύει κείμενα μόνο στην ελληνική γλώσσα.
- Ο καθαρισμός κειμένου, ουσιαστικά η διαδικασία προεπεξεργασίας του κειμένου που συμβάλλει στη μετέπειτα ποιοτικότερη ανάλυση του, αντιμετωπίζει κάποια προβλήματα. Ενδεικτικό είναι πως κείμενα που περιέχουν μαθηματικούς τύπους δεν αναλύονται επιτυχώς, οπότε ο χρήστης καλείται να τους εντοπίσει και να τους απομακρύνει ο ίδιος. Μια διαδικασία που ίσως φαντάζει εύκολη, όμως σε ένα μακροσκελές κείμενο με αρκετούς μαθηματικούς τύπους στο περιεχόμενο του, σίγουρα θα βοηθούσε η αυτοματοποίηση της διαδικασίας αυτής.
- Ορισμένες φορές η εφαρμογή μπορεί να χρειαστεί αρκετό χρόνο ώστε να πραγματοποιήσει τις απαιτούμενες διεργασίες και να εμφανίσει στο χρήστη τα αποτελέσματα που αιτήθηκε να λάβει.
- Η εφαρμογή έχει σχεδιαστεί για να δέχεται κείμενα με επέκταση .txt, .pdf και .docx. Η εξαγωγή κειμένου από τέτοια αρχεία δεν είναι πάντοτε επιτυχής (κυρίως τα αρχεία .pdf και .docx αντιμετωπίζουν θέμα). Ο χρήστης καλείται να ελέγξει το κείμενο μετά της εισαγωγή του στην εφαρμογή και σε περίπτωση που αποτύχει η αποκωδικοποίηση του να το εισάγει χειροκίνητα.

5.Προτάσεις Βελτίωσης - Συμπεράσματα

5.1 Προτάσεις Βελτίωσης

Ο ‘Texter’ είναι μια εφαρμογή που δεν πρωτοπορεί σε κάποιο κομμάτι έναντι άλλων αντίστοιχων εφαρμογών ή εργαλείων που προϋπήρχαν. Ωστόσο, σε αυτό το σημείο θα παραθέσω ορισμένες βελτιώσεις που με τη μελλοντική προσθήκη τους θα μπορούσαν να μετατρέψουν τον ‘Texter’ σε ένα εργαλείο ιδιαίτερα χρήσιμο τόσο για ερευνητές όσο και για επαγγελματίες διαφόρων κλάδων.

1. ΣΥΝΕΡΓΑΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ: ένα ιδιαίτερα σημαντικό χαρακτηριστικό που συνοδεύει όλες τις πρωτοπόρες εφαρμογές του κλάδου. Η δυνατότητα να διαμοιράζονται και να κοινοποιούνται αρχεία και αποτελέσματα μεταξύ των μελών της εφαρμογής όπως επίσης και οι υποδομές για επικοινωνία και συνεργασία μεταξύ των χρηστών είναι στοιχεία που αναβαθμίζουν την εφαρμογή και μπορούν να φανούν βοηθητικά στη διεκπεραίωση ομαδικών εργασιών και στο διαμοιρασμό καθηκόντων.
2. ΕΞΑΓΩΓΗ ΚΕΙΜΕΝΟΥ: όπως ειπώθηκε και παραπάνω η εφαρμογή συναντάει προβλήματα στην εξαγωγή κειμένου από ορισμένους τύπους αρχείων. Η ποιοτικότερη αποκωδικοποίηση του περιεχομένου των αρχείων κειμένου καθώς επίσης και η εξαγωγή κειμένου με προηγμένες τεχνικές (π.χ. εξαγωγή κειμένου από εικόνες) είναι ένα ακόμα χαρακτηριστικό που με την προσθήκη του θα αναβάθμιζε την εικόνα της εφαρμογής.
3. ΟΠΤΙΚΟΠΟΙΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ: αυτή τη στιγμή δεν προσφέρεται καμία δυνατότητα οπτικοποίησης των αποτελεσμάτων. Η προσθήκη ειδικών διαγραμμάτων οπτικοποίησης των αποτελεσμάτων της ανάλυσης όπως π.χ. η χρήση word cloud, word trees κ.α. μπορούν να επιταχύνουν και να βελτιώσουν την κατανόηση των πληροφοριών που μας παρέχει η ανάλυση και να οδηγήσουν σε ποιοτικότερα συμπεράσματα ωστόσο χρειάζεται προσοχή στον τρόπο αναπαράστασης των εξαγόμενων πληροφοριών καθώς μια σύνθετη και πολύπλοκη απεικόνιση τους θα μπορούσε να οδηγήσει σε παρερμηνείες. Υπάρχουν κάποιες δοκιμασμένες τεχνικές παρόλα αυτά ο συγκεκριμένος τομέας βρίσκεται υπό συνεχή ανάπτυξη και βελτίωση με στόχο την πληρέστερη αναπαράσταση αποτελεσμάτων χωρίς αυτή να ελλοχεύει κινδύνους για παρερμηνείες.

4. ΟΜΑΔΟΠΟΙΗΣΗ - ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ: στην ενότητα ‘Τα κείμενα μου’ θα μπορούσε να προστεθεί η δυνατότητα δημιουργίας φακέλων για να ομαδοποιούνται έγγραφα με παρόμοιο περιεχόμενο είτε αυτοματοποιημένα είτε με την παρέμβαση του χρήστη.

5.2 Συμπεράσματα

Σε αυτή την εργασία προσπαθήσαμε να ξεκαθαρίσουμε τι είναι η ανάλυση κειμένου, που βρίσκει εφαρμογή και πως πραγματοποιείται, εξετάζοντας τις τεχνικές και τα στάδια που λαμβάνουν μέρος κατά τη διαδικασία της ανάλυσης.

Όπως έγινε αντιληπτό πρόκειται για ένα θέμα ‘ομπρέλα’ κάτω από το οποίο συναντάμε μια πληθώρα επιστημονικών πεδίων όπως αυτά της μηχανικής μάθησης, της τεχνητής νοημοσύνης, της στατιστικής, της ανάκτησης πληροφοριών, της επεξεργασίας φυσικής γλώσσας και της εξόρυξης πληροφοριών. Αν και η ανάλυση κειμένου ξεκίνησε με στόχο να βοηθήσει τους ερευνητές να μειώσουν το χρόνο που σπαταλούσαν για την εύρεση πληροφοριών μέσω της απλής ευρετηριοποίησης όρων, είδαμε ότι εξελίχθηκε σε κάτι πολύ μεγαλύτερο. Σήμερα βρίσκει εφαρμογή σε πλήθος κλάδων ανάμεσα στους οποίους η υγεία, η ασφάλεια, η εξυπηρέτηση πελατών, η ανάλυση μέσων κοινωνικής δικτύωσης, το μάρκετινγκ, η διοίκηση κ.α.

Αρκεί να αναλογιστούμε πως περίπου το 80% των πληροφοριών μιας επιχείρησης υπολογίζεται ότι περιέχονται σε έγγραφα κειμένου. Είναι επομένως υψηλής εμπορικής και στρατηγικής σημασίας η ανάκτηση, η διαχείριση και η εξαγωγή χρήσιμων πληροφοριών από τα κειμενικά έγγραφα που έχει στην κατοχή της η εκάστοτε εταιρεία ή οργανισμός.

Όσο η τεχνολογία εξελίσσεται τόσο μικραίνει το κόστος συντήρησης των δεδομένων. Όμως θα ήταν ανούσιο να έχει κανείς στην κατοχή του τόσα δεδομένα χωρίς να τα αξιοποιεί. Έτσι μαζί με τη ραγδαία εξέλιξη της τεχνολογίας ακολουθεί και η ραγδαία εξέλιξη των επιστημονικών πεδίων που προαναφέραμε και με την εξέλιξη αυτή έρχονται και οι προκλήσεις. Προκλήσεις που η επιστημονική κοινότητα καλείται να ξεπεράσει, με τις κυριότερες να εντοπίζονται στο στάδιο της προεπεξεργασίας των δεδομένων αυτών και στην προσπάθεια ‘καθαρισμού’ τους και δόμησης τους. Σημαντικά εμπόδια επίσης θεωρούνται η πολυγλωσσικότητα που χαρακτηρίζει τα κειμενικά δεδομένα όπως και η οπτικοποίηση τους, μιας και ο σχεδιασμός τέτοιων διεπαφών είναι ιδιαίτερα ευαίσθητο ζήτημα.

Παράλληλα με την προσπάθεια που έγινε να κατανοήσουμε τι είναι η ανάλυση κειμένου, πως λειτουργεί και που χρησιμεύει δημιουργήθηκε και η διαδικτυακή εφαρμογή ‘Texter’ που συνοδεύει την παρούσα εργασία. Πρόκειται για μια εφαρμογή που περιορίζεται στην εξαγωγή συμπερασμάτων σε ποσοτικό επίπεδο μιας και οι αναλύσεις που πραγματοποιεί αφορούν κυρίως αριθμητικά αποτελέσματα, όμως αυτό δεν εμποδίζει τους χρήστες από το να καταλήξουν οι ίδιοι σε συμπεράσματα ποιοτικού περιεχομένου.

Καταλήγοντας, πιστεύω είναι εμφανές πως η πολυκλαδική επιστήμη της ανάλυσης κειμένων λειτουργεί καταλυτικά στην ανακάλυψη γνώσης και όπως είπε ο Φράνσις Μπέικον η “Γνώση είναι δύναμη”.

Βιβλιογραφία

https://www.ceu.edu/tanad/what

https://guides.lib.fsu.edu/text-analysis/definitions

https://www.postgresql.org/docs/current/textsearch-controls.html

Acerbi A., Lampos V., Garnett P. & Bentley A. (2013 March 20). The expression of emotions in 20th century books. PLOS ONE. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059030

Akilan A. (2015). Text mining: Challenges and future directions. 2015, 2nd International Conference on Electronics and Communication Systems (ICECS), 1679–1684. https://doi.org/10.1109/ECS.2015.7124872

Bail C. (2012). The fringe effect: Civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. American Sociological Review, 77 (6), 855 – 879.

Berelson B. (1952). Content analysis in communication research. Glencoe IL: Free Press.

Colley S. K. & Neal A. (2012). Automated text analysis to examine qualitative differences in safety schema among upper managers supervisors and workers. Safety Science, 50 (9), 1775 – 1785.

Deepika Sharma, Stemming Algorithms A Comparative Study and their Analysis, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA, Volume 4 – No.3, September 2012 – www.ijais.org.

Dianna R. Mullet. 'A General Critical Discourse Analysis Framework for Educational Research' (February 20, 2018).

Dr. S. Vijayarani et al. International Journal of Computer Science & Communication Networks, Vol 5(1), 7 - 16, (February 2015).

Eshbaugh - Soha M. (2010). The tone of local presidential news coverage. Political Communication, 27(2), 121 – 140.

Evison J. (2013). Turn openings in academic talk: Where goals and roles intersect. Classroom Discourse, 4(1), 3 – 26.
Fairclough N. (1995). Critical discourse analysis: The critical study of language. London England: Longman.

Fayyad U. M. Piatetsky - Shapiro G. & Smyth P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In Knowledge Discovery and Data Mining (pp. 82–88).

Foucault M. (1973). The order of things: An archaeology of the human sciences. New York, NY: Vintage Books.

Frakes W. B. & Baeza - Yates R. (1992). Information Retrieval: Data Structures & Algorithms. New Jersey: Prentice Hall.

G. Salton and C. Buckley "Term-Weighting Approaches in Automatic Text Retrieval" Information Processing and Management: An Int'l J., vol. 24, no. 5, pp. 513 - 523, 1988.

Gabe Ignatow, Rada Mihalcea.'An Introduction to Text Mining Research Design, Data Collection and Analysis' (2018).

Grimmer J. & Stewart B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21 (3), 267 – 297.

H. Ahonen, O. Heinonen, M. Klemettinen and A.I. Verkamo "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections" Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2 - 11, 1998.

Hearst M. (1999). Untangling text data mining. In Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics.

Heritage J. & Raymond G. (2005). The terms of agreement: Indexing epistemic authority and subordination in talk - in - interaction. Social Psychology Quarterly, 68 (1), 15 – 38.

Hotho A. Nürnberger A. and Paaß G. (2005). "A brief survey of text mining". In Ldv Forum, Vol. 20 (1), p. 19 - 62.

Kumar V. & Joshi M. (2003). What is data mining?

Lazard A., Scheinfeld E., Bernhardt J., Wilcox G. & Suran M. (2015). Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. American Journal of Infection Control, 43 (10), 1109 – 1111.

Leskovec J. Grobelnik M. & Milic-Frayling N. (2004). Learning sub-structures of document semantic graphs for document summarization. In KDD 2004 Workshop on Link Analysis and Group Detection (LinkKDD) Seattle Washington.

Levenberg A., Pulman S., Moilanen K., Simpson E. & Roberts S. (2014). Predicting economic indicators from web text using sentiment composition. Retrieved from http://www.robots.ox.ac.uk/~parg/pubs/sentiment_ICICA2014.pdf

Marwick B. (2013). Discovery of emergent issues and controversies in anthropology using text mining, topic modeling and social network analysis of microblog content. In C. Yonghua & Y. Zhao (Eds.), Data mining applications with R, 63 – 93. Cambridge, England: Academic Press.

Menaka S and Radha N, Text Classification using Keyword Extraction Technique, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013, ISSN: 2277 128X.

Mische A. (2014). Measuring futures in action: Projective grammars in the Rio + 20 debates. Theory & Society 43 (3 – 4) 437 – 464.

Mitchell T. (1997). Machine Learning. McGraw-Hill.

Mohan V, (2015). Preprocessing Techniques for Text Mining — An Overview.

Ms. Anjali Ganesh Jivani (2011) A Comparative Study of Stemming Algorithms, Anjali Ganesh Jivani et al., Int. J. Comp. Tech. Appl., Vol 2 (6), 1930 - 1938, ISSN: 2229 - 6093.

Parker I. (1992). Discourse dynamics: Critical analysis for social and individual psychology. London, England: Routledge.

Raja U., Mitchell T., Day T. & Hardin J. (2008). Text mining in healthcare. Applications and opportunities. Journal of healthcare information management : JHIM, 22, 52 – 56.

Roberts C. W. (1997). Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts. Mahwah, NJ: Lawrence Erlbaum.

Rockwell Geoffrey. 'What Is Text Analysis Really?' Literary and Linguistic Computing 18, τχ. 2 (Ιούνιος, 2003): 209 – 219.

S. Shehata, F. Karray and M. Kamel "A Concept-Based Model for Enhancing Text Categorization", Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629 - 637, 2007.

S. Shehata, F. Karray and M. Kamel "Enhancing Text Clustering Using Concept-Based Mining Model", Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043 - 1048, 2006.

S.Charanyaa and K.Sangeetha, Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization, International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 2, February 2014, ISSN (Online): 2320 - 9801.

S.Jusoh and H.M. Alfawareh, Natural language interface for online sales, in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007). Malaysia: IEEE, November 2007, pp. 224 – 228.

Schmitt R. (2005). Systematic metaphor analysis as a method of qualitative research. The Qualitative Report, 10 (2), 358 – 394.

VijayGaikwad S., Chaugule A. & Patil P. (2014). Text Mining Methods and Techniques. International Journal of Computer Applications, 85 (17), 42 – 45. https://doi.org/10.5120/14937-3507

Yanli Cai, Jian - Tao Sun: Text Mining. Encyclopedia of Database Systems (2nd ed.) 2018.