ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

‘‘Μελέτη αλγορίθμων εξόρυξης σημαντικών λέξεων/φράσεων από κείμενο και δημιουργία εφαρμογής’’

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του

ΑΜΑΤΕΟ ΑΛΙΟΥ

(4255)

Επιβλέπων : **Νίκος Δημόκας**

Επίκουρος Καθηγητής

Καστοριά 7 Ιουλίου – 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

‘‘Μελέτη αλγορίθμων εξόρυξης σημαντικών λέξεων/φράσεων από κείμενο και δημιουργία εφαρμογής’’

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του

ΑΜΑΤΕΟ ΑΛΙΟΥ

(4255)

Επιβλέπων : **Νίκος Δημόκας**

Επίκουρος Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1/7/2024

.....	.....	.....
Νίκος Δημόκας Επίκουρος Καθηγητής	Δημήτριος Ι. Βέργαδος, Πρόεδρος του Τμήματος, Αναπληρωτής Καθηγητής	Ιωάννης Βαρδάκας Αναπληρωτής Καθηγητής

Καστοριά 7 Ιουλίου – 2024

Ευχαριστίες

Θεωρώ υποχρέωση μου να ευχαριστήσω τον επιβλέποντα καθηγητή για την πολύτιμη καθοδήγησή του. Επιπλέον, αισθάνομαι την ανάγκη να αφιερώσω αυτή την εργασία στους κοντινούς μου ανθρώπους, οι οποίοι με στήριξαν αδιάλειπτα καθ' όλη τη διάρκεια των σπουδών μου στο Πανεπιστήμιο Δυτικής Μακεδονίας. Η αγάπη και η συμπαράστασή τους υπήρξαν καθοριστικοί παράγοντες στην επιτυχία μου.

Περίληψη

Η παρούσα εργασία πραγματεύεται την αυτόματη εξαγωγή λέξεων και φράσεων-κλειδιών από κείμενα. Στα πλαίσια της εργασίας, μελετήθηκε διεξοδικά η ερευνητική περιοχή της εξόρυξης κειμένου και πιο συγκεκριμένα, εξετάστηκαν αλγόριθμοι εξαγωγής λέξεων-κλειδιών όπως οι RAKE, YAKE, TF-IDF και TextRank. Παράλληλα, η εργασία επεκτείνει τον υπάρχοντα αλγόριθμο RAKE, προσθέτοντας σημασιολογία στον τρόπο με τον οποίο εξάγονται οι λέξεις-κλειδιά. Στόχος της παρούσας εργασίας είναι η επαύξηση του αλγορίθμου RAKE για την πιο αποδοτική και αποτελεσματική εξαγωγή λέξεων-κλειδιών, οι οποίες να αντιπροσωπεύουν με τον καλύτερο δυνατό τρόπο το κείμενο από το οποίο εξάγονται.

Λέξεις Κλειδιά: Εξόρυξη Κειμένου, Εξόρυξη Δεδομένων, Λέξεις-Κλειδιά, RAKE, YAKE, TF-IDF, TextRank, Σημασιολογία

Abstract

This paper deals with the automatic extraction of key words and phrases from texts. As part of the work, the research area of text mining was thoroughly studied and more specifically, keyword extraction algorithms such as RAKE, YAKE, TF-IDF and TextRank were examined. In parallel, the work extends the existing RAKE algorithm by adding semantics to the way keywords are extracted. The aim of this work is to augment the RAKE algorithm for the most efficient and effective extraction of keywords, which represent in the best possible way the text from which they are extracted.

Key Words: Text Mining, Data mining, Keyword Extraction, RAKE, YAKE, TF-IDF, TextRank, Semantics

Πίνακας Περιεχομένων

Λίστα Εικόνων

Λίστα Πινάκων

Εισαγωγή

Η εξόρυξη κειμένου είναι μια τεχνολογία που επιτρέπει την ανάλυση μεγάλων κειμένων για την εξαγωγή χρήσιμων πληροφοριών και γνώσεων. Χρησιμοποιεί τεχνικές από τομείς όπως η μηχανική μάθηση, η επεξεργασία φυσικής γλώσσας και η ανάλυση δεδομένων για να εντοπίσει μοτίβα και σχέσεις μέσα στα κείμενα. Η εξόρυξη κειμένου μπορεί να εφαρμοστεί σε διάφορα πεδία, όπως η ανάλυση συναισθημάτων σε μέσα κοινωνικής δικτύωσης, η αναγνώριση θεμάτων σε επιστημονικά άρθρα, και η ανάλυση πελατειακών σχολίων για τη βελτίωση προϊόντων και υπηρεσιών. Η διαδικασία αυτή περιλαμβάνει στάδια όπως η προ επεξεργασία του κειμένου (π.χ. αφαίρεση κοινών λέξεων και ορθογραφικές διορθώσεις), η ανάλυση και η εξαγωγή των δεδομένων, και η ερμηνεία των αποτελεσμάτων, προσφέροντας πολύτιμες γνώσεις που θα ήταν δύσκολο να εντοπιστούν χειροκίνητα.

Η εξαγωγή λέξεων-κλειδιών είναι μια διαδικασία στην οποία συγκεκριμένες λέξεις ή φράσεις εντοπίζονται και εξάγονται από ένα κείμενο επειδή θεωρούνται ότι αντιπροσωπεύουν με ακρίβεια το περιεχόμενο και τα κύρια θέματα του. Αυτή η διαδικασία είναι σημαντική για την ανάλυση και την κατηγοριοποίηση κειμένων, τη βελτίωση της αναζήτησης πληροφοριών και την κατανόηση μεγάλων συνόλων δεδομένων κειμένου. Οι λέξεις-κλειδιά μπορούν να εξάγονται χρησιμοποιώντας διάφορους αλγορίθμους και τεχνικές, όπως ο TF-IDF, ο RAKE, ο YAKE, και ο TextRank. Αυτές οι τεχνικές χρησιμοποιούν διαφορετικές μεθόδους για να εντοπίσουν τις πιο σημαντικές λέξεις, λαμβάνοντας υπόψη παραμέτρους όπως η συχνότητα εμφάνισης, η θέση των λέξεων στο κείμενο και η συνάφεια με το γενικό πλαίσιο του κειμένου.

Ο στόχος της παρούσας εργασίας είναι η δημιουργία ενός αλγορίθμου που θα επιτρέπει την πιο αποδοτική και αποτελεσματική εξαγωγή λέξεων-κλειδιών από κείμενα, προσθέτοντας σημασιολογία στον υπάρχοντα αλγόριθμο RAKE. Η εργασία επικεντρώνεται στη διεξοδική μελέτη της ερευνητικής περιοχής της εξόρυξης κειμένου και των αλγορίθμων εξαγωγής λέξεων-κλειδιών, όπως ο RAKE, ο YAKE, ο TF-IDF, και ο TextRank Μέσα από αυτήν την μελέτη, επιδιώκεται η βελτίωση της ακρίβειας και της αποδοτικότητας του αλγορίθμου RAKE, ώστε να αντιπροσωπεύει καλύτερα το περιεχόμενο των κειμένων. Το τελικό αποτέλεσμα θα είναι ένα λογισμικό που θα μπορεί να χρησιμοποιηθεί σε διάφορες εφαρμογές, από την ανάλυση επιστημονικών άρθρων μέχρι την αξιολόγηση πελατειακών σχολίων, προσφέροντας μια πιο ακριβή και λεπτομερή κατανόηση του κειμένου.

1. Εξόρυξη Δεδομένων

Σήμερα, υπάρχει μια πληθώρα όρων που χρησιμοποιούνται για να περιγράψουν αυτήν τη διαδικασία. Οι όροι αυτοί περιλαμβάνουν τα αναλυτικά στοιχεία, τις προγνωστικές αναλύσεις, τα μεγάλα δεδομένα, τη μηχανική μάθηση και την ανακάλυψη γνώσης σε βάσεις δεδομένων. Παρ' όλα αυτά, όλοι αυτοί οι όροι έχουν ένα κοινό στόχο: την εξαγωγή γνώσης από μεγάλα σύνολα δεδομένων. Έτσι, χρησιμοποιούμε τον όρο “εξόρυξη δεδομένων” για να αναφερθούμε σε αυτήν τη διαδικασία [1, 2].

1.1 Αρχές Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων [7, 11], ως ονομαζόμενη προσπάθεια, εμφανίστηκε στα τέλη της δεκαετίας του 1980 από την κοινότητα της βάσης δεδομένων, η οποία αναρωτιόταν από πού θα προέρχονταν τα επόμενα μεγάλα βήματα προς τα εμπρός. Με την ανάπτυξη και την επιτυχημένη εφαρμογή της θεωρίας της σχεσιακής βάσης δεδομένων, άνοιξε ο δρόμος για τη συλλογή μεγάλων ποσοτήτων δεδομένων. Πώς μπορούμε να αξιοποιήσουμε αυτά τα τεράστια αποθέματα δεδομένων για να προσθέσουμε αξία;

Οι πρώτες προσπάθειες εξόρυξης δεδομένων στις αρχές της δεκαετίας του 1990 προσέλκυσαν ερευνητές από την κοινότητα των βάσεων δεδομένων. Σύντομα, και άλλες επιστήμες της πληροφορικής, κυρίως η τεχνητή νοημοσύνη, κέρδισαν επίσης ενδιαφέρον. Σημειωτέο είναι ότι ένα βασικό στοιχείο της “νοημοσύνης” είναι η ικανότητα μάθησης. Εδώ και πολλά χρόνια, η έρευνα στη μηχανική μάθηση έχει αναπτύξει τεχνολογίες που αντιμετωπίζουν αυτήν την πρόκληση. Η μηχανική μάθηση αφορά τη συλλογή δεδομένων παρατήρησης μέσω της αλληλεπίδρασης με τον κόσμο και τη δημιουργία μοντέλων του κόσμου από αυτά τα δεδομένα. Αυτό συνάδει σε μεγάλο βαθμό με τον σκοπό της εξόρυξης δεδομένων, και έτσι οι κοινότητες της μηχανικής μάθησης και της εξόρυξης δεδομένων συνεργάζονται όλο και περισσότερο [13].

Παρόλο που η στατιστική είναι ένα από τα θεμελιώδη εργαλεία για την ανάλυση δεδομένων, έχει ξεπεράσει τα εκατό χρόνια ιστορίας. Τα στατιστικά στοιχεία φέρνουν στο τραπέζι βασικές ιδέες σχετικά με την αβεβαιότητα και τον τρόπο με τον οποίο μπορούμε να την προσαρμόσουμε στα μοντέλα που κατασκευάζουμε. Οι στατιστικές παρέχουν ένα πλαίσιο για την αξιολόγηση της ακρίβειας και της αξιοπιστίας των μοντέλων που δημιουργούμε από τα δεδομένα. Οι ανακαλύψεις πρέπει να είναι στατιστικά έγκυρες και σημαντικές, και κάθε αβεβαιότητα που σχετίζεται με τη μοντελοποίηση πρέπει να εξετάζεται προσεκτικά. Τα στατιστικά δεδομένα διαδραματίζουν κεντρικό ρόλο στη σύγχρονη εξόρυξη δεδομένων.

Σήμερα, η εξόρυξη δεδομένων αποτελεί έναν κλάδο που βασίζεται σε προηγμένες δεξιότητες στην επιστήμη των υπολογιστών, τη μηχανική μάθηση και τη στατιστική.

1.2 Τι είναι Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων είναι η εξεύρεση, καθαρισμός, επεξεργασία και εξαγωγή σημαντικής, μη τετριμμένης, προηγούμενα άγνωστης και πιθανά χρήσιμης πληροφορίας ή προτύπων από δεδομένα που βρίσκονται αποθηκευμένα σε μεγάλα σύνολα δεδομένων [8-10]. Αυτή η διαδικασία περιλαμβάνει συνήθως την εφαρμογή τεχνικών και αλγορίθμων από τον τομέα της στατιστικής, της τεχνητής νοημοσύνης και της μηχανικής μάθησης σε μεγάλα σύνολα δεδομένων για να ανακαλυφθούν κρυμμένα μοτίβα ή συσχετίσεις. Υπάρχει μεγάλη ποικιλία όσον αφορά τους τομείς προβλημάτων, τις εφαρμογές, τις διατυπώσεις και τις αναπαραστάσεις δεδομένων που συναντώνται σε πραγματικές εφαρμογές όπως για παράδειγμα επιστημονικές έρευνες και ιατρικές εφαρμογές. Έτσι, η “εξόρυξη δεδομένων” αναφέρεται σε μια ευρεία γκάμα τεχνικών που χρησιμοποιούνται για την ανάλυση και την εξαγωγή πληροφοριών από μεγάλα σύνολα δεδομένων. Ο πιο διαδεδομένος ορισμός της “εξόρυξης δεδομένων” είναι η διαδικασία ανακάλυψης “μοντέλων” δεδομένων. Όμως πολλά διαφορετικά στοιχεία αποτελούν ένα μοντέλο [4-6].

Σήμερα, υπάρχει μια ευρεία γκάμα όρων που χρησιμοποιούνται για να περιγράψουν αυτήν τη διαδικασία, συμπεριλαμβανομένων της ανάλυσης δεδομένων, της προγνωστικής ανάλυσης, των “μεγάλων δεδομένων” (big data), της μηχανικής μάθησης και της ανακάλυψης γνώσης σε βάσεις δεδομένων. Ωστόσο, όλοι αυτοί οι όροι έχουν ως κοινό στόχο την ανάκτηση ενεργών ψηφίδων γνώσης από μεγάλα σύνολα δεδομένων. Γι' αυτό το λόγο, ο όρος “εξόρυξη δεδομένων” χρησιμοποιείται σε αυτή την διαδικασία για να αναπαρασταθεί σε ολόκληρο το κείμενο [4].

1.3 Χρησιμότητα Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων παρέχει αξία βοηθώντας στο να κατανοήσει κάνεις καλυτέρα πως λειτουργεί μια επιχείρηση δηλαδή πως σχετίζονται τα στοιχεία μιας επιχείρησης μεταξύ τους. Παρέχει πληροφορίες σχετικά με τις ενέργειες που μπορεί να ακολουθήσει μια επιχείρηση προκειμένου να βελτιώσει τη λειτουργία της και να αυξήσει τα έσοδά της. Επίσης μπορεί να βοηθήσει να προσδιοριστεί που μπορεί να μειώσει το κόστος χωρίς να βλάψει τον οργανισμό και που οι δαπάνες αποφέρουν τις καλύτερες αποδόσεις [6, 7, 12, 13].

Παρέχει μια πληθώρα οφελών για τις επιχειρήσεις και τους επιστήμονες δεδομένων. Ας ρίξουμε μια ματιά σε μερικά από τα κυριότερα πλεονεκτήματά [1, 11, 14] της:

- Ανακάλυψη Κρυφών Πληροφοριών: Η εξόρυξη δεδομένων επιτρέπει την εντοπισμό κρυφών πληροφοριών και μοτίβων, ανοίγοντας νέους ορίζοντες για την καινοτομία και την ανάπτυξη.

- Λήψη Αποφάσεων Βασισμένη σε Δεδομένα: Η δυνατότητα λήψης αποφάσεων βασισμένη σε δεδομένα επιτρέπει στις επιχειρήσεις να είναι πιο ευέλικτες και αποτελεσματικές στις δράσεις τους.
- Αύξηση Αποδοτικότητας και Αποτελεσματικότητας: Η βελτίωση των διαδικασιών και η αύξηση της αποδοτικότητας είναι κρίσιμες για την επιτυχία της επιχείρησης, και η εξόρυξη δεδομένων βοηθά σε αυτό τον τομέα.
- Πρόβλεψη Συμπεριφοράς: Η δυνατότητα πρόβλεψης μελλοντικών συμπεριφορών επιτρέπει στις επιχειρήσεις να προσαρμόζονται προληπτικά στις αλλαγές της αγοράς.
- Βελτίωση Αποφάσεων: Η διαθεσιμότητα αξιόπιστων δεδομένων βοηθά τις επιχειρήσεις να λαμβάνουν πιο ενημερωμένες και ακριβείς αποφάσεις.
- Αύξηση Ασφάλειας: Η εξόρυξη δεδομένων βοηθά στον εντοπισμό πιθανών απειλών και κινδύνων, προστατεύοντας έτσι τις επιχειρήσεις από απάτες και απώλειες.

Παρά τα πολλά πλεονεκτήματα, υπάρχουν και ορισμένα μειονεκτήματα [12, 13, 21] που πρέπει να ληφθούν υπόψη:

- Ανάγκη για Σωστά Δεδομένα: Η εξόρυξη δεδομένων απαιτεί υψηλής ποιότητας δεδομένα για να παράγει αξιόπιστα αποτελέσματα. Λανθασμένα ή ατελή δεδομένα μπορούν να οδηγήσουν σε λανθασμένες αναλύσεις και συμπεράσματα.

- Πολυπλοκότητα Εφαρμογής: Η εξόρυξη δεδομένων απαιτεί τη χρήση εξειδικευμένων εργαλείων και γνώσεων. Η πολυπλοκότητα αυτή μπορεί να αποθαρρύνει ορισμένες επιχειρήσεις ή οργανισμούς από την υιοθέτηση της εξόρυξης δεδομένων.

- Ανάγκη Προστασίας της Ιδιωτικότητας: Η χρήση προσωπικών δεδομένων στην εξόρυξη δεδομένων μπορεί να προκαλέσει ανησυχίες σχετικά με την ιδιωτικότητα και την ασφάλεια των δεδομένων.

- Κίνδυνος Overfitting (υπερπροσαρμογή): Υπάρχει ο κίνδυνος να δημιουργηθούν μοντέλα που είναι υπερβολικά προσαρμοσμένα στα δεδομένα εκπαίδευσης και να μην γενικεύονται καλά σε νέα δεδομένα.

1.4 Τι περιλαμβάνει η Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων συνδυάζει ιδέες από τις βάσεις δεδομένων, την στατιστική, την τεχνητή νοημοσύνη, την μηχανική μάθηση και την αναγνώριση προτύπων [1, 4, 6]. Χρησιμοποιεί τις βάσεις δεδομένων ως πηγή δεδομένων για την ανάλυση και εξαγωγή των πληροφοριών. Πάνω σε αυτά χρησιμοποιεί την στατιστική για να ανάλυση τα δεδομένα, να εκτίμηση πιθανότητες και για να επεξεργάζεται αβέβαια δεδομένα. Την τεχνητή νοημοσύνη την χρησιμοποιεί επειδή παρέχει αλγορίθμους και τεχνικές για την

ανάλυση και εξαγωγή συμπερασμάτων από τα δεδομένα. Η μηχανική μάθηση προσφέρει τεχνικές για την ανάπτυξη μοντέλων και την εκπαίδευση τους από τα δεδομένα ώστε να προβλέπουν μελλοντικά στοιχεία η να ανακαλύπτουν μοτίβα. Τέλος η αναγνώριση προτύπων χρησιμοποιείται για να αναγνωρίζει συγκεκριμένα δεδομένα ή μοτίβα εντός μια μεγάλης συλλογής δεδομένων [10, 12, 13].

1.5 Εφαρμογές Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων αναδεικνύεται ως αναγκαία διαδικασία λόγω της ασταμάτητης αύξησης του όγκου των διαθέσιμων δεδομένων [6, 10, 12]. Η εύκολη συλλογή και διαθεσιμότητα των δεδομένων μαζί με τα εργαλεία αυτοματοποιημένης συλλογής και τις πολλές πήγες μεγάλου όγκου δεδομένων, από επιστήμες όπως είναι η ιατρική, βιολογία, φυσική και οικονομία, από επιχείρησής όπως το e-commerce και τα οικονομικά δεδομένα και από κοινωνικά όπως το web και τα μέσα κοινωνικής δικτύωσης, έχουν αυξήσει δραματικά τη διαθεσιμότητα των δεδομένων. Προσφέρει αυτοματοποιημένες τεχνικές ανάλυσης μεγάλου όγκου δεδομένων επιτρέποντας την ανακάλυψη κρυμμένων πληροφοριών και μοτίβων που διαφεύγουν την αντίληψη με παραδοσιακές μεθόδους. Τέλος η ανάπτυξη της εξόρυξης δεδομένων οφείλεται επίσης στην ανάγκη για ανακάλυψη νέων συσχετίσεων ή προτύπων στα δεδομένα που μπορούν να παρέχουν νέες ευκαιρίες, προβλέψεις η κατανόηση συμπεριφορών και τάσεων. Έτσι αποτελεί κρίσιμη διαδικασία για την αποκάλυψη των ανεξερεύνητων ενδεχομένων και την αντιμετώπιση της αυξανομένης πολυπλοκότητας των δεδομένων στη σύγχρονη κοινωνικά.

Χρησιμοποιείται σε πολλούς τομείς όπως οι επιχειρήσεις, η ιατρική, η εκπαίδευση, η επιστήμη κλπ. Μερικά παραδείγματα εφαρμογής είναι η πρόβλεψη των πωλήσεων, αναγνώριση απάτης, πρόβλεψη νοσημάτων, προσωποποιημένες συστάσεις και ανάλυση κοινωνικών μέσων [9, 10, 11, 14]. Πιο αναλυτικά:

- Μάρκετινγκ και πωλήσεις: Η εξόρυξη δεδομένων βοηθά τις εταιρείες να κατανοήσουν τη συμπεριφορά των καταναλωτών, τις προτιμήσεις και τα μοτίβα αγορών. Επιτρέπει στοχευμένες καμπάνιες μάρκετινγκ, εξατομικευμένες προτάσεις και πρόβλεψη πωλήσεων αναγνωρίζοντας μοτίβα στις πωλήσεις όπως τα προϊόντα που πωλούνται συχνότερα μαζί η τα προϊόντα που πωλούνται καλυτέρα σε συγκεκριμένες περιοχές η εποχές.
- Χρηματοοικονομικά: Στα χρηματοοικονομικά, η εξόρυξη δεδομένων χρησιμοποιείται για βαθμολόγηση πιστώσεων, ανίχνευση απάτης, διαχείριση κινδύνου και αλγοριθμικές συναλλαγές. Βοηθά τα χρηματοπιστωτικά ιδρύματα να λαμβάνουν τεκμηριωμένες αποφάσεις και να εντοπίζουν πιθανούς κινδύνους.
- Διαχείριση Σχέσεων Πελατών: Αναλύοντας τα δεδομένα των πελατών, οι επιχειρήσεις μπορούν να βελτιώσουν την ικανοποίηση, τη διατήρηση και την

πιστότητα των πελατών τους. Η εξόρυξη δεδομένων βοηθά στην τμηματοποίηση των πελατών, στην πρόβλεψη της απόκλισης και στη βελτιστοποίηση των στρατηγικών μάρκετινγκ.

- Κατασκευή και διαχείριση αλυσίδας εφοδιασμού: Η εξόρυξη δεδομένων βελτιστοποιεί τις διαδικασίες παραγωγής, προβλέπει αστοχίες εξοπλισμού και βελτιώνει τη διαχείριση αποθεμάτων. Βοηθά τους οργανισμούς να εξορθολογήσουν τις λειτουργίες και να μειώσουν το κόστος.
- Τηλεπικοινωνίες: Οι εταιρείες τηλεπικοινωνιών χρησιμοποιούν την εξόρυξη δεδομένων για βελτιστοποίηση δικτύου, πρόβλεψη απόκλισης πελατών και στοχευμένο μάρκετινγκ. Βοηθά στη βελτίωση της ποιότητας των υπηρεσιών και στην ικανοποίηση των πελατών.
- Ανίχνευση απάτης και κυβερνοασφάλεια: Οι τεχνικές εξόρυξης δεδομένων είναι ζωτικής σημασίας για τον εντοπισμό δόλιων δραστηριοτήτων σε χρηματοοικονομικές συναλλαγές, εντοπισμό ανωμαλιών στα μοτίβα χρήσης των πελατών που ενδεχομένως να υποδηλώνουν απατή, ασφαλιστικές αξιώσεις και διαδικτυακές συναλλαγές. Βοηθά επίσης στον εντοπισμό απειλών για την ασφάλεια και στην προστασία ευαίσθητων πληροφοριών.
- Υγειονομική περίθαλψη: Η ανάλυση δεδομένων ασθενών βελτιώνει την ποιότητα της υγειονομικής περίθαλψης, επιτρέποντας τον εντοπισμό των τάσεων της νόσου, την πρόβλεψη των αποτελεσμάτων των ασθενών και την εξατομίκευση των θεραπευτικών προσεγγίσεων. Βοηθά επίσης στην ιατρική έρευνα, στην ανακάλυψη φαρμάκων και μπορεί να χρησιμοποιηθεί για να προβλέψει πιθανά νοσήματα σε μια πληθυσμιακή ομάδα, βασισμένη σε παράγοντες όπως ιατρικό ιστορικό και γενετικά δεδομένα.
- Προσωποποιημένες συστάσεις: Πολλές επιχειρήσεις χρησιμοποιούν την εξόρυξη δεδομένων για να δώσουν προσωποποιημένες συστάσεις στους πελάτες τους, όπως συστάσεις για προϊόντα ή υπηρεσίες βασισμένες στις προηγούμενες αγοραστικές τους συνήθειες.
- Ανάλυση κοινωνικών μέσων: Οι εταιρείες μπορούν να χρησιμοποιήσουν την εξόρυξη δεδομένων για να αναλύσουν δεδομένα από κοινωνικά μέσα, όπως Twitter και Facebook, για να κατανοήσουν την αντίδραση του κοινού σε συγκεκριμένα προϊόντα ή γεγονότα. Η ακόμα και για την ανάλυση συναισθήματος, ανίχνευση τάσεων και στοχευμένη διαφήμιση. Βοηθά τις επιχειρήσεις να κατανοήσουν τις απόψεις των πελατών και τις τάσεις της αγοράς
- Εκπαίδευση: Στον τομέα της εκπαίδευσης, η εξόρυξη δεδομένων βοηθά στην ανάλυση της απόδοσης των μαθητών, στη σύσταση μαθημάτων και στον εντοπισμό παραγόντων που επηρεάζουν τη διατήρηση των μαθητών. Βοηθά τους εκπαιδευτικούς να βελτιώσουν τις μεθόδους διδασκαλίας και τα εκπαιδευτικά αποτελέσματα.

1.6 Διαδικασία Εξόρυξης Δεδομένων

Η διαδικασία εξόρυξης δεδομένων ακολουθεί συγκεκριμένα στάδια ξεκινώντας με την συλλογή των αρχικών δεδομένων που θα χρησιμοποιηθούν για την ανάλυση και ορίζοντας το πρόβλημα που θέλουμε να επιλύσουμε με σαφήνεια [2, 6]. Έπειτα εξασφαλίζουμε τους απαραίτητους πόρους συμπεριλαμβάνοντας το ανθρώπινο δυναμικό, υλικό και λογισμικό. Αφαιρούνται πιθανά σφάλματα, ατελή δηλαδή να υπάρχει έλλειψη τιμών κάποιων χαρακτηριστικών, αντιφάσεις, ασυνεπή δηλαδή να έχουν διαφορές σε ονόματα, κωδικούς, ημερομηνίες και ανωμαλίες από τα δεδομένα και υπόκεινται σε διάφορες διαδικασίες επεξεργασίες όπως κανονικοποίηση, μείωση της διάστασης, επιλογή χαρακτηριστικών κτλ. και στην συνέχεια εφαρμόζονται οι αλγόριθμοι εξόρυξης δεδομένων για να ανακαλυφθούν μοτίβα, τάσεις και συσχετίσεις στα δεδομένα. Όταν ανακαλυφθούν αξιολογούνται τα αποτελέσματα ως προς την ακρίβεια και την χρησιμότητα τους και ερμηνεύονται και αναλύονται προκειμένου να εξαχθούν συμπεράσματα. Τα αποτελέσματα και οι γνώσεις που αποκτήθηκαν χρησιμοποιούνται για την λήψη αποφάσεων η την υποστήριξη σε συγκεκριμένα προβλήματα. Η γνώση που αποκτήθηκε μπορεί να χρησιμοποιηθεί για να καταστεί η λήψη αποφάσεων πιο αποτελεσματική και η επίλυση μελλοντικών προβλημάτων πιο αποτελεσματική και αποδοτική [1, 10, 11, 12]. Αυτά μπορούν να ενσωματωθούν στην καθημερινή λειτουργία μιας επιχείρησης η μιας οργάνωσης για τη βελτίωση της απόδοσης η την ανάπτυξη καινοτόμων λύσεων [13].

1.7 Διαδικασίες προεπεξεργασίας

Όπως αναφέρθηκε προηγουμένως, η διαδικασία εξόρυξης δεδομένων είναι ένας αγωγός που περιέχει πολλές φάσεις. Σε αυτή την ενότητα, θα μελετήσουμε αυτές τις διαφορετικές φάσεις [11, 12].

1.7.1 Καθορισμός Δεδομένων

Το στάδιο καθορισμού δεδομένων (data cleansing) [2, 11, 13] είναι σημαντικό λόγω των σφαλμάτων που σχετίζονται με την διαδικασία εξόρυξης δεδομένων. Αναλαμβάνει την διόρθωση και τη βελτίωση της ποιότητας των δεδομένων πριν την εφαρμογή αλγορίθμων εξόρυξης. Αυτό περιλαμβάνει την ανακάλυψη και την αντιμετώπιση κελιών με ελλιπής τιμές, με την αντικατάσταση τους με μέση τιμή η με άλλες κατάλληλες τιμές ανάλογα με το πεδίο και τον σκοπό τους. Άλλες τεχνικές που εφαρμόζονται είναι για την ανίχνευση και την αφαίρεση θορύβου προκειμένου να διατηρηθεί η ακρίβεια και η αξιοπιστία τους. Επίσης εντοπίζονται και αφαιρούνται τιμές που μπορεί να επηρεάζουν

αρνητικά την ανάλυση η τα αποτελέσματα και αντιμετωπίζονται πιθανές ασυνέπειες όπως αντιφάσεις η αντικείμενα που δεν συμφωνούν με τα προκαθορισμένα κριτήρια.

1.7.2 Μετασχηματισμός δεδομένων

Ο μετασχηματισμός δεδομένων [11, 13] είναι ένα στάδιο στην διαδικασία της εξόρυξης δεδομένων κατά το οποίο τα δεδομένα υποβάλλονται σε διάφορες τεχνικές επεξεργασίας για να είναι πιο κατάλληλα για την εφαρμογή των αλγορίθμων εξόρυξης. Υπάρχουν 2 μετασχηματισμοί δεδομένων που μπορούν να γίνουν. Αυτές είναι η κανονικοποίηση και διακριτοποίηση δεδομένων [1, 2]. Πιο αναλυτικά:

- Κανονικοποίηση δεδομένων: Η κανονικοποίηση αφορά τη μετατροπή των τιμών των δεδομένων σε ένα συγκεκριμένο εύρος ή κλίμακα, συνήθως με σκοπό να εξαλειφθεί η επίδραση των διαφορών στις κλίμακες των χαρακτηριστικών στην ανάλυση. Η κανονικοποίηση μπορεί να γίνει με τη χρήση τεχνικών όπως η min-max κλίμακα ή η z-score κανονικοποίηση.
- Διακριτοποίηση δεδομένων: Κατά τη διακριτοποίηση τα συνεχή χαρακτηριστικά μετατρέπονται σε διακριτά δηλαδή χωρίζονται σε διακριτές κατηγορίες ή διαστήματα τιμών. Αυτό μπορεί να βοηθήσει στην αντιμετώπιση προβλημάτων που απαιτούν τη χρήση διακριτών δεδομένων ή στη μείωση της πολυπλοκότητας των μοντέλων.

1.7.3 Μεταβολή όγκου δεδομένων

Η μεταβολή του όγκου [10, 11, 13] των δεδομένων συνήθως περιλαμβάνει τη μείωση της διάστασης των δεδομένων και/ή τον περιορισμό του αριθμού των δεδομένων που χρησιμοποιούνται για την ανάλυση. Και οι δύο αυτές προσεγγίσεις βοηθούν στη μείωση του όγκου των δεδομένων, ενώ ταυτόχρονα διατηρούν ή βελτιώνουν την αξιοπιστία και την απόδοσή τους κατά τη διαδικασία της εξόρυξης δεδομένων. Παρακάτω τα εξετάζουμε αναλυτικά.

- Μείωση διάστασης δεδομένων (επιλογή χαρακτηριστικών): Επιλέγονται τα πιο σημαντικά χαρακτηριστικά από το σύνολο των δεδομένων. Αυτό μπορεί να γίνει με τη χρήση τεχνικών όπως η ανάλυση δεσμών (cluster analysis) ή η ανάλυση κύριων συνιστωσών (principal component analysis - PCA), που επιτρέπουν την αναγνώριση των πιο επιδραστικών χαρακτηριστικών.
- Μείωση αριθμού δεδομένων: Επιλέγεται μόνο ένα υποσύνολο των δεδομένων για ανάλυση. Αυτό μπορεί να είναι χρήσιμο όταν τα αρχικά δεδομένα είναι υπερβολικά μεγάλα ή όταν έχουμε ένα πολύ μεγάλο αριθμό δειγμάτων. Η επιλογή προτύπων μπορεί να γίνει με βάση κριτήρια όπως η αναπαράσταση της διαφορετικότητας των δεδομένων ή η εξάλειψη περιττής πληροφορίας.

1.8 Τεχνικές και μέθοδοι Εξόρυξης Δεδομένων

Ανάλογα με τα χαρακτηριστικά των δεδομένων και τους στόχους της ανάλυσης χρησιμοποιούνται και οι ανάλογες τεχνικές και μέθοδοι. Κάποιες από αυτές είναι η κατηγοριοποίηση, η παλινδρόμηση, η ανάλυση χρονοσειρών, η ομαδοποίηση, κανόνες συσχέτισης και ανάλυση ακολουθιών.

1.8.1 Κατηγοριοποίηση

Η κατηγοριοποίηση [2, 3, 6, 8, 13] αναφέρεται στη διαδικασία τοποθέτησης ενός νέου στοιχείου σε μια κατηγορία, από ένα προκαθορισμένο σύνολο βάσει των χαρακτηριστικών του. Ο στόχος της κατηγοριοποίησης είναι η ανάπτυξη ενός μοντέλου που θα πραγματοποιεί κατηγοριοποίηση σε μελλοντικών άγνωστα δεδομένα δηλαδή ο στόχος είναι η εκπαίδευση ενός μοντέλου ή ενός αλγορίθμου ώστε να μπορεί να προβλέπει την κατηγορία ενός νέου στοιχείου με βάση τα χαρακτηριστικά που του έχουν δοθεί. Ένα παράδειγμα εφαρμογής της κατηγοριοποίησης είναι η αυτόματη ταξινόμηση email σε φακέλους όπως “Εισερχόμενα”, “Διαφήμιση” η “Σημαντικά”. Αυτό επιτυγχάνεται με την εκπαίδευση ενός μοντέλου μηχανικής μάθησης με χαρακτηριστικά όπως ο τίτλος, ο περιεχόμενος, ο αποστολέας κ.λπ., έτσι ώστε να αποφασίζει αυτόματα σε ποιον φάκελο ανήκει ένα νέο email. Οι αλγόριθμοι κατηγοριοποίησης μπορεί να είναι απλοί, όπως ο αλγόριθμος k-Nearest Neighbors ο οποίος αξιοποιεί την έννοια της απόστασης μεταξύ των δεδομένων για την ταξινόμηση. Στην ουσία, όταν πρέπει να ταξινομήσουμε ένα νέο σημείο δεδομένων, το KNN επιλέγει τα k πλησιέστερα σημεία δεδομένων στον χώρο χαρακτηριστικών και βασίζεται στην κλάση την οποία εκπροσωπούν αυτά τα σημεία για να καθορίσει την κλάση του νέου σημείου. Η πιο σύνθετοι, όπως τα Νευρωνικά Δίκτυα τα οποία αναπτύσσονται για την ανάλυση πολύπλοκων δεδομένων και προβλημάτων. Αυτά τα μοντέλα αποτελούνται από πολλά επίπεδα νευρώνων που λειτουργούν όπως το ανθρώπινο εγκέφαλο. Η εκπαίδευση τους γίνεται με την παρουσίαση μεγάλου όγκου δεδομένων και την προσαρμογή των βαρών τους ώστε να εκτελούν τις επιθυμητές λειτουργίες, όπως η αναγνώριση προτύπων ή η πρόβλεψη τιμών. Είναι ιδιαίτερα αποτελεσματικά σε προβλήματα που περιλαμβάνουν μεγάλο όγκο δεδομένων ή πολυπλοκότητα, όπως η εξόρυξη συναισθήματος ή η αναγνώριση προτύπων σε εικόνες. Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από τους στόχους της κατηγοριοποίησης και τον τύπο των δεδομένων. Το κύριο πλεονέκτημα της κατηγοριοποίησης είναι η δυνατότητα αυτόματης ταξινόμησης των δεδομένων σε κατηγορίες, που επιτρέπει την εξαγωγή συμπερασμάτων και την λήψη αποφάσεων βάσει τους.

1.8.2 Παλινδρόμηση

Στην παλινδρόμηση [3, 6, 9] προσπαθούμε να καταλάβουμε τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών, όπως ανεξάρτητες μεταβλητές (π.χ. χαρακτηριστικά) και εξαρτημένες μεταβλητές (π.χ. απόκριση ή πρόβλεψη), μέσω μιας συνάρτησης που περιγράφει τη σχέση ανάμεσά τους. Είναι κυρίως χρήσιμη όταν θέλουμε να προβλέψουμε μια συνεχή απόκριση (εξαρτημένη μεταβλητή) βάσει μιας ή περισσότερων μεταβλητών (ανεξάρτητων μεταβλητών). Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε τη μέθοδο της παλινδρόμησης για να προβλέψουμε την τιμή ενός ακινήτου, λαμβάνοντας υπόψη χαρακτηριστικά όπως η τοποθεσία, το μέγεθος και άλλες σχετικές παράμετροι. Οι πιο δημοφιλείς αλγόριθμοι παλινδρόμησης περιλαμβάνουν τη γραμμική παλινδρόμηση και τη λογαριθμική παλινδρόμηση. Στη γραμμική παλινδρόμηση, προσπαθούμε να προσαρμόσουμε μία ευθεία στα δεδομένα, επιχειρούμε να προσαρμόσουμε μία γραμμή στα δεδομένα, έτσι ώστε να αντιπροσωπεύει τη σχέση μεταξύ των μεταβλητών. Στην ουσία, προσπαθούμε να βρούμε την καλύτερη γραμμή που ταιριάζει με τα δεδομένα μας. Ο στόχος η μείωση της απόκλισης μεταξύ των πραγματικών και των προβλεπόμενων τιμών της εξαρτημένης μεταβλητής. Ενώ στη λογαριθμική παλινδρόμηση προσπαθούμε να προσαρμόσουμε μία καμπύλη όπου οι μεταβλητές μετασχηματίζονται με τον φυσικό λογάριθμο πριν από την εφαρμογή της παλινδρόμησης. Αυτός ο μετασχηματισμός χρησιμοποιείται συχνά όταν οι δεδομένες τιμές έχουν τάση να αυξάνονται ή να μειώνονται εκθετικά. Η λογαριθμική παλινδρόμηση μπορεί να είναι πιο κατάλληλη από τη γραμμική όταν οι σχέσεις μεταξύ των μεταβλητών δεν είναι γραμμικές στη φυσική τους μορφή, αλλά γίνονται γραμμικές μετά από το μετασχηματισμό. Αυτή η παραλλαγή μπορεί να βελτιώσει την προσαρμογή του μοντέλου στα δεδομένα και να παράγει πιο ακριβείς προβλέψεις. Η παλινδρόμηση είναι μία ισχυρή τεχνική για την ανάλυση και πρόβλεψη συναρτήσεων από δεδομένα, και χρησιμοποιείται ευρέως σε πολλούς τομείς, συμπεριλαμβανομένων των επιστημών των υπολογιστών, της οικονομίας, της βιολογίας και πολλών άλλων. Η επιτυχής εφαρμογή της παλινδρόμησης απαιτεί την κατανόηση των δεδομένων και την επιλογή του κατάλληλου μοντέλου παλινδρόμησης για το συγκεκριμένο πρόβλημα.

1.8.3 Ανάλυση χρονοσειρών

Η ανάλυση χρονοσειρών [5, 8, 13] είναι μία εξειδικευμένη προσέγγιση η οποία επικεντρώνεται στην κατανόηση, την πρόβλεψη και την ανάλυση δεδομένων που παρουσιάζουν χρονική εξέλιξη. Αυτό μπορεί να περιλαμβάνει χρονοσειρές όπως οι οικονομικές δείκτες, ο καιρός, η κίνηση σε μια πόλη και πολλά άλλα. Οι βασικές διαδικασίες στην ανάλυση χρονοσειρών περιλαμβάνουν τα:

- Αναγνώριση προτύπων: Η αναγνώριση προτύπων περιλαμβάνει την εντοπισμό τυχόν τάσεων, εποχιακών παραλλαγών και άλλων προτύπων στη χρονοσειρά. Η

αναγνώριση αυτών των προτύπων είναι σημαντική για την κατανόηση της δυναμικής των δεδομένων.

- Πρόβλεψη: Στόχος της πρόβλεψης είναι να προβλεφθεί η μελλοντική τιμή της χρονοσειράς βάσει των παρατηρήσεων του παρελθόντος. Αυτό μπορεί να επιτευχθεί με τη χρήση μοντέλων πρόβλεψης όπως τα αυτοσυνεχιζόμενα μοντέλα (autoregressive models), τα μοντέλα εξισορρόπησης (smoothing models), ή τεχνικές όπως η μέθοδος Holt-Winters.

- Διαγνωστική ανάλυση: Η διαδικασία της διαγνωστικής ανάλυσης εστιάζει στην εκτίμηση της ποιότητας των προβλέψεων και στην αναγνώριση τυχόν προβληματικών περιοχών ή ανωμαλιών στη χρονοσειρά.

- Χρονοσειριακή συσταδοποίηση: Η χρονοσειριακή συσταδοποίηση σημαίνει την ομαδοποίηση χρονοσειρών σε ομάδες με παρόμοιες χρονοσειριακές τάσεις ή πρότυπα, προκειμένου να ανακαλυφθούν κοινά χαρακτηριστικά.

Η ανάλυση χρονοσειρών είναι κρίσιμη σε πολλούς τομείς όπως η οικονομία, η υγεία, η κλιματολογία και η παραγωγή, καθώς βοηθά στην κατανόηση των τάσεων και των μοτίβων που κρύβονται στα δεδομένα με χρονική διάσταση, καθώς και στη λήψη αποφάσεων βάσει αυτών των αναλύσεων.

1.8.4 Ομαδοποίηση

Η ομαδοποίηση [1, 2, 3, 8, 13], γνωστή επίσης και ως clustering, είναι μία διαδικασία που στοχεύει στο να χωρίσει ένα σύνολο δεδομένων σε ομάδες ή συστάδες, ώστε τα στοιχεία που ανήκουν σε μία ομάδα να είναι πιο ομοιόμορφα μεταξύ τους σε σχέση με στοιχεία σε άλλες ομάδες. Είναι μια αποτελεσματική τεχνική όταν δεν υπάρχουν προκαθορισμένες ετικέτες και κατηγορίες για τα δεδομένα, και όταν επιθυμούμε να εντοπίσουμε φυσικά σύνολα ή πρότυπα στα δεδομένα χωρίς προκαθορισμένη γνώση για τη δομή τους. Κατά την ομαδοποίηση, είναι σημαντικό να οριστεί πώς θα μετρηθεί η ομοιότητα μεταξύ των στοιχείων. Αυτό μπορεί να γίνει με τη χρήση διαφόρων μετρικών όπως η ευκλείδεια απόσταση, η απόσταση Manhattan, η συνημιτονική ομοιότητα κ.λπ. Μία πρόκληση στην ομαδοποίηση είναι η επιλογή του κατάλληλου αριθμού ομάδων. Υπάρχουν διάφορες μετρικές και τεχνικές που μπορούν να χρησιμοποιηθούν για αυτό τον σκοπό, όπως ο δείκτης Silhouette και ο αλγόριθμος Elbow. Η ομαδοποίηση έχει εφαρμογές σε πολλούς τομείς, όπως η ανάλυση κοινωνικών δικτύων, η ανάλυση πελατών σε επιχειρήσεις, η αναγνώριση μοτίβων σε χρονοσειρές, και πολλά άλλα. Η επιτυχής ομαδοποίηση μπορεί να οδηγήσει σε καλύτερη κατανόηση των δεδομένων και τη διαμόρφωση πιο αποτελεσματικών στρατηγικών λήψης αποφάσεων. Υπάρχουν διάφορες μέθοδοι ομαδοποίησης, όπως οι ιεραρχικές μέθοδοι (hierarchical clustering), οι μέθοδοι βάσης κέντρων (centroid-based methods) και οι

μέθοδοι βάσης περιοχής (density-based methods). Οι ιεραρχικές μέθοδοι διαιρούν το σύνολο δεδομένων με βάση την απόσταση μεταξύ των στοιχείων και τις συστάδες που δημιουργούνται δομούνται ιεραρχικά. Οι μέθοδοι βάσης κέντρων χρησιμοποιούν έναν αριθμό κέντρων για να αντιπροσωπεύσουν κάθε συστάδα και ανατίθενται τα στοιχεία στο κοντινότερο κέντρο. Οι μέθοδοι βάσης περιοχής εντοπίζουν συστάδες με βάση την πυκνότητα των δεδομένων και είναι ευαίσθητες στις παραμέτρους, μπορώντας να ανακαλύψουν συστάδες με πολύπλοκες μορφές και μεγέθη.

1.8.5 Κανόνες συσχέτισης

Οι κανόνες συσχέτισης [2, 8, 9, 15] αποτελούν ένα εργαλείο στον χώρο της εξόρυξης δεδομένων και της ανάλυσης συσχετίσεων μεταξύ διαφορετικών στοιχείων ή γεγονότων σε ένα σύνολο δεδομένων. Αυτοί οι κανόνες εκφράζουν πιθανές συνδέσεις ή συσχετίσεις μεταξύ διαφορετικών γνωρισμάτων ή γεγονότων σε ένα σύνολο δεδομένων. Εκφράζονται συνήθως σε μορφή “Εάν ... τότε ...”, όπου το “Εάν” αναφέρεται σε μια συνθήκη ή ένα σύνολο συνθηκών που πρέπει να ικανοποιούνται, ενώ το “Τότε” αναφέρεται σε μια συνέπεια ή μια ενέργεια που συνδέεται με την ικανοποίηση αυτών των συνθηκών. Ένα παράδειγμα είναι εάν ο πελάτης αγοράζει αυγά και γάλα, τότε πιθανόν να αγοράσει και ψωμί. Οι κανόνες συσχέτισης είναι χρήσιμοι για την ανίχνευση συνδέσεων και συμπεριφορών μεταξύ διαφορετικών γεγονότων ή στοιχείων σε ένα σύνολο δεδομένων. Εφαρμόζονται συχνά σε προβλήματα συσταδοποίησης, πρόβλεψης συνεπειών και διαμόρφωσης στρατηγικών πωλήσεων. Μπορεί να παρέχει στιγμιότυπα για τις συνήθειες, τις προτιμήσεις και τις τάσεις των χρηστών ή των πελατών και να βοηθήσει στη λήψη αποφάσεων βασισμένων σε αυτές τις ανακαλύψεις.

1.8.6 Ανάλυση ακολουθιών

Η ανάλυση ακολουθιών [2, 11, 13] επικεντρώνεται στην εξαγωγή πληροφοριών από δεδομένα που περιέχουν ανακάλυψη προτύπων, τάσεων, συσχετίσεων σε ακολουθίες δεδομένων, σειρές ή ακολουθίες διακριτών ή συνεχών στοιχείων. Αυτό μπορεί να περιλαμβάνει ακολουθίες γεγονότων, όπως η συμπεριφορά χρηστών σε μια ιστοσελίδα, το κείμενο λόγου ή ακόμα και χρονοσειρές, κείμενα, γενετικοί κώδικες, ακολουθίες γεγονότων κ.λπ. Οι τεχνικές που χρησιμοποιούνται για την ανάλυση ακολουθιών περιλαμβάνουν την εξόρυξη προτύπων, την πρόβλεψη και την κατηγοριοποίηση. Όπου στην εξόρυξη προτύπων (Pattern Mining), αναζητούμε ενδιαφέρουσες και σημαντικές συνεκτικές ακολουθίες μέσα στα δεδομένα. Αυτό μπορεί να είναι η ανίχνευση συχνών συμβάντων, αναγνώριση ακολουθιών γεγονότων με συγκεκριμένη σημασία ή ανίχνευση αρνητικών προτύπων. Στην πρόβλεψη ακολουθιών (Sequence Prediction), προσπαθούμε να προβλέψουμε την επόμενη τιμή ή

το επόμενο γεγονός σε μια ακολουθία. Αυτό μπορεί να γίνει με τη χρήση μοντέλων όπως τα αναδρομικά νευρωνικά δίκτυα (RNNs) ή τα μοντέλα Markov. Τέλος στην κατηγοριοποίηση ακολουθιών (Sequence Classification), αναθέτουμε μια κατηγορία ή ετικέτα σε μια ακολουθία βάσει των παρατηρήσεων της. Αυτό μπορεί να γίνει με τη χρήση μεθόδων όπως οι αναδρομικοί ταξινομητές (Recursive Classifiers) ή με μεθόδους όπως το κρυμμένο μοντέλο του Μαρκόβ (Hidden Markov Models).

2. Εξόρυξη Κείμενου

Σήμερα, σχεδόν όλες οι υπάρχουσες πληροφορίες σε διαφορετικά ιδρύματα (π.χ. κυβέρνηση, επιχειρήσεις, βιομηχανία και άλλα) διατηρούνται σε ηλεκτρονικά έγγραφα στα οποία περιέχουν ημιδομημένα δεδομένα. Σε ένα έγγραφο, η “περίληψη” αποτελεί ένα μη δομημένο στοιχείο κειμένου. Αντίθετα, δομημένα πεδία σε ένα έγγραφο περιλαμβάνουν το όνομα του συγγραφέα, την ημερομηνία δημοσίευσης, τον τίτλο και την κατηγορία [17]. Σύμφωνα με μια μελέτη από [18], η εξόρυξη κειμένου έχει καταστεί ένα από τα σύγχρονα πεδία που έχει ενσωματωθεί σε πολλούς ερευνητικούς τομείς, όπως η υπολογιστική γλωσσολογία, η ανάκτηση πληροφοριών και η εξόρυξη δεδομένων. Η εξόρυξη κειμένου διαφέρει από την εξόρυξη δεδομένων. Ενώ η εξόρυξη δεδομένων [19] επικεντρώνεται στην ανακάλυψη ενδιαφέροντων μοτίβων από μεγάλες βάσεις δεδομένων, η εξόρυξη κειμένου αναζητά πληροφορίες εντός του κειμένου [20].

Μεθοδολογίες ανάκτησης πληροφοριών, όπως τεχνικές δημιουργίας ευρετηρίου κειμένου, έχουν αναπτυχθεί για το χειρισμό μη δομημένων εγγράφων. Σε συμβατικές έρευνες, θεωρείται ότι ένας χρήστης αναζητά κυρίως γνωστούς όρους, που έχουν χρησιμοποιηθεί ή γραφτεί στο παρελθόν από κάποιον άλλο. Το κύριο πρόβλημα είναι ότι τα αποτελέσματα αναζήτησης δεν είναι σχετικά με τις απαιτήσεις του χρήστη. Μια λύση είναι να χρησιμοποιήσετε την εξόρυξη κειμένου για να βρείτε σχετικές πληροφορίες, οι οποίες δεν αναφέρονται ρητά ούτε έχουν σημειωθεί μέχρι στιγμής. Η διαδικασία της εξόρυξης κειμένου ξεκινά με τη συλλογή εγγράφων μέσω διαφορετικών πόρων. Ένα συγκεκριμένο έγγραφο θα ανακτηθεί μέσω του οργάνου εξόρυξης κειμένου και ελέγχοντας τη μορφή και τα σύνολα χαρακτήρων του θα υποβληθεί σε προεπεξεργασία από αυτό το μέσο. Αμέσως μετά, το έγγραφο υποβάλλεται σε ανάλυση κειμένου. Αυτή η διαδικασία περιλαμβάνει τη σημασιολογική ανάλυση με σκοπό την απόκτηση υψηλής ποιότητας πληροφοριών μέσω του κειμένου. Υπάρχουν πολλές μέθοδοι ανάλυσης κειμένου που μπορούν να χρησιμοποιηθούν ανάλογα με τον στόχο του οργανισμού. Σε ορισμένες περιπτώσεις, η ανάλυση του κειμένου επαναλαμβάνεται πολλές φορές μέχρι να εξαχθούν οι απαραίτητες πληροφορίες. Τα αποτελέσματα μπορούν να αποθηκευτούν σε ένα σύστημα διαχείρισης πληροφοριών που παρέχει έναν μεγάλο όγκο σημαντικών πληροφοριών για τον χρήστη αυτού του συστήματος. Ένα συγκεκριμένο έγγραφο ανακτάται μέσω του μηχανισμού εξόρυξης κειμένου, και υποβάλλεται σε προεπεξεργασία από τον ίδιο τον μηχανισμό.

Η ιστορία [23, 24, 25] της εξόρυξης κειμένου είναι στενά συνδεδεμένη με την εξέλιξη της πληροφορικής και της τεχνητής νοημοσύνης [22]. Οι πρώτες μορφές εξόρυξης κειμένου αναπτύχθηκαν στη δεκαετία του 1980, αλλά η πραγματική έκρηξη στην έρευνα και την εφαρμογή της εξόρυξης κειμένου συνέβη στα τέλη της δεκαετίας του 1990 και τις αρχές του 21ου αιώνα. Η ανάπτυξη των μεθόδων και τεχνικών που

χρησιμοποιούνται σήμερα στην εξόρυξη κειμένου έγινε δυνατή λόγω της συνεχούς εξέλιξης της υπολογιστικής ισχύος και της ανάπτυξης νέων αλγορίθμων και τεχνικών στον τομέα της τεχνητής νοημοσύνης και της επεξεργασίας φυσικής γλώσσας. Αυτή η εξέλιξη έχει επιτρέψει την ανάπτυξη πολύπλοκων συστημάτων εξόρυξης κειμένου που μπορούν να αναλύσουν και να εξάγουν πληροφορίες από μεγάλους όγκους δεδομένων από κείμενα [26]. Σημαντικές στιγμές στην ιστορία της εξόρυξης κειμένου περιλαμβάνουν:

- Αρχές της Δεκαετίας του 1980: Οι πρώτες προσπάθειες εξόρυξης κειμένου εμφανίστηκαν με τη χρήση απλών στατιστικών μεθόδων για την ανάλυση και την ταξινόμηση κειμένων.

- Τέλη της Δεκαετίας του 1990 - Αρχές του 21ου Αιώνα: Η έρευνα στον τομέα της εξόρυξης κειμένου εκτοξεύτηκε με την ανάπτυξη προηγμένων μοντέλων επεξεργασίας φυσικής γλώσσας και την εφαρμογή τεχνικών μηχανικής μάθησης στον τομέα.

- Τα Τελευταία Χρόνια: Η εξόρυξη κειμένου έχει γίνει κρίσιμη σε πολλούς τομείς, όπως η επιχειρηματική ανάλυση, η υγεία, η κοινωνική επιστήμη και άλλοι, με συνεχή εξέλιξη και καινοτομία στον τομέα.

Η εξόρυξη κειμένου [27, 28, 29, 30] είναι μια διαδικασία υπολογιστικής ανάλυσης και εξαγωγής πληροφοριών από κείμενα [31]. Στόχος της είναι να ανακαλύψει μοτίβα, τάσεις, συναισθήματα, θέματα ή οποιαδήποτε άλλη χρήσιμη πληροφορία από μεγάλα σύνολα κειμένων. Για να επιτευχθεί αυτό, χρησιμοποιούνται διάφορες τεχνικές όπως η φυσική γλώσσα επεξεργασία (Natural Language Processing-NLP) [27], η μηχανική μάθηση, η τεχνητή νοημοσύνη και άλλες μέθοδοι υπολογιστικής επεξεργασίας κειμένου [28]. Μέσω αυτών των τεχνικών, το σύστημα μπορεί να αναγνωρίσει διαφορετικά στοιχεία όπως λέξεις-κλειδιά, οντότητες, θέματα, σχέσεις μεταξύ λέξεων και συμβάλει στην κατανόηση του περιεχομένου και των παραμέτρων που επηρεάζουν την ποιότητα και την περιεκτικότητα του κειμένου.

2.1 Χρησιμότητα και εφαρμογές Εξόρυξης Κειμένου

Η χρησιμότητα της εξόρυξης κειμένου είναι προφανής στην καθημερινή ζωή και τις επιχειρήσεις [29]. Με την εξέλιξη των μέσων κοινωνικής δικτύωσης και την αυξανόμενη δημοσιότητα τους, οι οργανισμοί αντιμετωπίζουν την πρόκληση να παρακολουθούν και να αναλύουν τις δημόσιες προσπάθειές τους. Ταυτόχρονα, η αύξηση του διαδικτυακού περιεχομένου και η ανάγκη για ψηφιοποίηση παλαιότερων εγγράφων δημιουργούν μια τεράστια ποσότητα κειμένου που πρέπει να διαχειριστούν. Οι νέες τεχνολογίες, όπως η αυτόματη μεταγραφή ήχου, βοηθούν στην αντιμετώπιση αυτής της πρόκλησης και στην αποθήκευση πολύτιμων δεδομένων. Έτσι, η εξόρυξη κειμένου αναδεικνύεται ως

κρίσιμο εργαλείο για την ανάλυση, την πρόβλεψη και τη λήψη αποφάσεων στις σύγχρονες επιχειρήσεις [33].

Ωστόσο, οι σύγχρονες τεχνολογικές εταιρείες εδράζονται κυρίως σε αριθμητικά και κατηγορικά δεδομένα για την απόκτηση πληροφοριών, τη χρήση αλγορίθμων μηχανικής μάθησης ή την εφαρμογή λειτουργικής βελτιστοποίησης [29]. Είναι αντιφατικό για μια επιχείρηση να επικεντρώνεται μόνο σε δομημένα δεδομένα, αγνοώντας τη σημασία της μη δομημένης φυσικής γλώσσας. Η ανάλυση του κειμένου αποτελεί μια ανεκμετάλλευτη πηγή πληροφοριών που μπορεί να δώσει σημαντικό ανταγωνιστικό πλεονέκτημα. Τέλος, οι επιχειρήσεις πλέον μεταβαίνουν από τη βιομηχανική εποχή στην εποχή της πληροφορίας. Οι πιο επιτυχημένες εταιρείες επαναπροσδιορίζουν την προσέγγισή τους και στρέφονται ξανά προς μια πελατοκεντρική φιλοσοφία. Αυτές οι εταιρείες κατανοούν ότι η μακροπρόθεσμη ευημερία των πελατών τους διασφαλίζει και τη δική τους μακροπρόθεσμη επιτυχία, διατηρώντας παράλληλα την ανταγωνιστικότητά τους. Οι μεγάλες εταιρείες δεν μπορούν πλέον να παράγουν απλώς ένα προϊόν και να το διαθέτουν στην αγορά χωρίς να λαμβάνουν υπόψη τις ανάγκες και τις επιθυμίες των τελικών χρηστών. Στη σημερινή εποχή, όπου οι προσδοκίες των πελατών αυξάνονται συνεχώς, οι πελάτες θέλουν να ακούν και να αισθάνονται ότι οι εταιρείες τους ακούν. Συνεπώς, για να διατηρήσουν έναν πραγματικά πελατοκεντρικό προσανατολισμό σε ένα ανταγωνιστικό περιβάλλον, οι εταιρείες πρέπει να είναι προσεκτικές και να ακούν τους πελάτες τους όσο το δυνατόν συχνότερα. Ωστόσο, ο όγκος των πληροφοριών που προκύπτει από αυτές τις αλληλεπιδράσεις μπορεί να είναι τεράστιος. Έτσι, η εξόρυξη κειμένου παρέχει έναν γρήγορο τρόπο εξαγωγής πληροφοριών. Επιτρέπει στους αναλυτές δεδομένων ή τους επιστήμονες να κατανοήσουν τεράστιες ποσότητες κειμένου και διασφαλίζει την αξιοπιστία των πληροφοριών από τους εσωτερικούς υπεύθυνους λήψης αποφάσεων [29]. Η μη χρήση εξόρυξης κειμένου μπορεί να σημαίνει αγνόηση πηγών κειμένου ή απλά δειγματοληψία και μη αυτόματη αναθεώρηση του κειμένου.

Η εξόρυξη κειμένου έχει εφαρμογές στην ανάλυση αισθημάτων, στα μέσα κοινωνικής δικτύωσης, της αυτόματης κατηγοριοποίησης κειμένων, της αναζήτησης πληροφοριών, επιχειρήσεις και οικονομία, υγειά, κοινωνικές επιστήμες, εκπαίδευση, ψηφιακή αρχειοθέτηση, νομική, πληροφορική, επιστήμη των δεδομένων και άλλα [27, 29, 32]. Πιο αναλυτικά:

- Στις επιχειρήσεις και στην οικονομία χρησιμοποιείται για την ανάλυση αναφορών, δελτίων τύπου, κειμένων στα social media, κριτικών πελατών και άλλων πηγών, προκειμένου να αναγνωρίσουν τάσεις, αντιλήψεις πελατών, και να λάβουν αποφάσεις βάσει αυτών των δεδομένων.

- Στην υγεία χρησιμοποιείται για την ανάλυση και την εξαγωγή πληροφορίας από ιατρικά άρθρα, κλινικές αναφορές, ιατρικές εγγραφές και άλλες πηγές,

προκειμένου να βοηθήσει στη διάγνωση, την αντιμετώπιση και την έρευνα σε ιατρικά ζητήματα.

- Στις κοινωνικές επιστήμες χρησιμοποιείται για την ανάλυση και την εξαγωγή συναισθημάτων, απόψεων, και τάσεων από κείμενα που αφορούν κοινωνικά θέματα, πολιτικά γεγονότα, κοινωνικά δίκτυα και άλλες πηγές.

- Στην εκπαίδευση χρησιμοποιείται για την ανάλυση εκπαιδευτικών υλικών, ερευνητικών άρθρων και αξιολόγησης φοιτητικών εργασιών, προκειμένου να αξιολογήσει την απόδοση των μαθητών και να βελτιώσει τις μεθόδους διδασκαλίας.

- Στην ψηφιακή αρχειοθέτηση χρησιμοποιείται για την οργάνωση, την αναζήτηση και την εξαγωγή πληροφοριών από ψηφιακά αρχεία κειμένου, όπως βιβλιοθήκες, μουσεία και άλλα πολιτιστικά ιδρύματα.

Η εξόρυξη κειμένου είναι σημαντική λόγου ότι βοηθά στην ανάκτηση πληροφοριών [27] από μεγάλους όγκους δεδομένων από κείμενα που θα ήταν αδύνατον ή εξαιρετικά χρονοβόρο να αναλυθούν με μηχανικά μέσα, επιτρέπει την ανίχνευση και την παρακολούθηση μοτίβων και τάσεων σε κείμενα, βοηθώντας στην πρόβλεψη μελλοντικών εξελίξεων και στη λήψη αποφάσεων, επιτρέπει την ανάλυση συναισθημάτων και απόψεων από κείμενα, που μπορεί να χρησιμοποιηθεί για την αξιολόγηση της δημόσιας γνώμης, την ανίχνευση δυνητικών προβλημάτων ή την ανάπτυξη στρατηγικών επικοινωνίας, βοηθά στη διαχείριση περιεχομένου, όπως στην αυτόματη κατηγοριοποίηση, σύνοψη και αναζήτηση πληροφοριών από μεγάλες βάσεις δεδομένων, παρέχει στήριξη στη λήψη αποφάσεων σε πολλούς τομείς, όπως στην επιχειρηματική στρατηγική, τη διαχείριση κινδύνων, την υγεία και την πολιτική, συμβάλλει στην ανάπτυξη νέων μεθόδων και εφαρμογών στους τομείς της τεχνητής νοημοσύνης, της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας. Συνολικά, η εξόρυξη κειμένου είναι σημαντική για την αξιοποίηση του πλούτου πληροφοριών που περιέχονται σε κείμενα, προκειμένου να ληφθούν πιο ενημερωμένες και ευστοχείς αποφάσεις σε διάφορους τομείς [33].

2.2 Διαδικασία

Η διαδικασία [27, 28, 31] της εξόρυξης κειμένου ξεκινά αρχικά συλλέγοντας τα κείμενα από τις κατάλληλες πήγες όπως ιστοσελίδες, βιβλία, άρθρα εφημερίδων, κείμενα από social media και άλλα. Αυτά τα κείμενα υπόκεινται σε επεξεργασία για την αφαίρεση περιττών στοιχείων όπως HTML tags, σημεία στίξης, ειδικούς χαρακτήρες και άλλα προβληματικά στοιχεία και διαχωρίζονται σε μικρότερες μονάδες κειμένου όπως προτάσεις η λέξεις για την επεξεργασία τους σε επόμενα στάδια και αναπαρίστανται σε μια μορφή που μπορεί να επεξεργαστεί από αλγορίθμους όπως διανύσματα λέξεων,

που επιτρέπουν τη σύγκριση και την ανάλυση τους. Έπειτα εφαρμόζονται τεχνικές εξόρυξης κειμένου, όπως η ανάλυση συναισθημάτων, η κατηγοριοποίηση, η οντολογική ανάλυση και άλλες, για την εξαγωγή πληροφοριών από τα κείμενα [33]. Τέλος, τα αποτελέσματα αξιολογούνται και ερμηνεύονται, για να εξαχθούν συμπεράσματα και να ληφθούν αποφάσεις βάσει της εξόρυξης του κειμένου. Κάθε ένα από αυτά τα βήματα απαιτεί προσεκτική επεξεργασία και τη χρήση κατάλληλων αλγορίθμων και τεχνικών, ανάλογα με τη φύση των δεδομένων και τους στόχους της εξόρυξης κειμένου.

2.3 Μέθοδοι Εξόρυξης Κείμενου

Οι μέθοδοι εξόρυξης κειμένου είναι ποικίλες και περιλαμβάνουν διάφορες τεχνικές και προσεγγίσεις για την ανάλυση και την εξαγωγή πληροφορίας από κείμενα. Ορισμένες από τις κύριες μεθόδους εξόρυξης κειμένου περιλαμβάνουν την κατηγοριοποίηση κειμένου (text classification), εξόρυξη γνώμης (opinion mining), εξαγωγή πληροφοριών (information extraction) και ανάλυση θεμάτων (analyzing topics). Πιο αναλυτικά:

2.3.1 Κατηγοριοποίηση κειμένου

Η κατηγοριοποίηση κειμένου (Text classification) [27, 28, 31] αποτελεί μια σημαντική εφαρμογή της τεχνητής νοημοσύνης που αφορά την ομαδοποίηση των κειμένων σε προκαθορισμένες κατηγορίες ή θεματικές ενότητες. Αυτό μπορεί να επιτευχθεί με τη χρήση διάφορων τεχνικών [32, 35], όπως η μηχανική μάθηση, οι αλγόριθμοι βαθιάς μάθησης και άλλες σχετικές μεθόδους. Πιθανότατα έχετε ήδη επωφεληθεί από τα οφέλη της κατηγοριοποίησης κειμένου χωρίς να το συνειδητοποιείτε. Για παράδειγμα, σκεφτείτε την επικοινωνία σας μέσω ηλεκτρονικού ταχυδρομείου. Περισσότερο από τα μισά από τα μηνύματα ηλεκτρονικού ταχυδρομείου που αποστέλλονται είναι ανεπιθύμητα, αλλά πιθανότατα τα εισερχόμενά σας δεν πλημμυρίζονται από ανεπιθύμητα μηνύματα. Αυτό οφείλεται στον κατηγοριοποίηση κειμένου [36] ανεπιθύμητης αλληλογραφίας που λειτουργεί στα παρασκήνια και διαχωρίζει τα νόμιμα μηνύματα από τα ανεπιθύμητα. Μερικά ακόμη παραδείγματα είναι:

- Αναγνώριση γλώσσας: Οι μηχανές αναζήτησης χρησιμοποιούν διάφορες τεχνικές ταξινόμησης προκειμένου να προσδιορίσουν τη γλώσσα του ερωτήματός σας. Αυτό το κάνουν για να μπορέσουν να κατευθύνουν το ερώτημά σας σε συλλογές ηλεκτρονικών εγγράφων που είναι γραμμένα στην ίδια γλώσσα.
- Ανάλυση συναισθήματος/Εξόρυξη απόψεων: Πρόκειται για την αξιολόγηση συναισθημάτων σε κείμενα, συχνά χρησιμοποιούμενη για να κατανοήσουμε τις απόψεις ή τα συναισθήματα των χρηστών για προϊόντα ή υπηρεσίες.

- Ανίχνευση Παραπλάνησης: Αυτή η εφαρμογή εστιάζει στον εντοπισμό ψευδών πληροφοριών ή εκφράσεων σε κείμενα, χρησιμοποιώντας γλωσσικά στοιχεία για τον εντοπισμό αναφορών που αποκαλύπτουν προσπάθειες παραπλάνησης.
- Άλλες Εφαρμογές: Υπάρχουν πολλές εφαρμογές ταξινόμησης κειμένων, περιλαμβάνοντας ταξινόμηση βάσει γλώσσας, είδους κειμένου, συναισθήματος, και περιεχομένου σε σχέση με τον αναγνώστη.

Η κατηγοριοποίηση κειμένων είναι μια εποπτευόμενη διαδικασία, καθώς απαιτεί την εκπαίδευση του συστήματος με ένα σύνολο κειμένων που έχουν ήδη ταξινομηθεί σε κατηγορίες. Αυτό το σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται για να μάθει το σύστημα τις συσχετίσεις μεταξύ των χαρακτηριστικών των κειμένων και των κατηγοριών τους. Η ομαδοποίηση κειμένων [22], από την άλλη πλευρά, είναι μια μη-εποπτευομένη διαδικασία όπου το σύστημα πρέπει να ομαδοποιήσει τα κείμενα χωρίς να έχει προηγούμενη γνώση των κατηγοριών. Αν και συχνά οι κατηγορίες μπορεί να είναι ιεραρχικές, η ομαδοποίηση κειμένων δεν απαιτεί προηγούμενη γνώση των κατηγοριών. Συνεπώς, η κατηγοριοποίηση κειμένου και η ομαδοποίηση κειμένου είναι διαφορετικές διαδικασίες με διαφορετικούς στόχους και μεθόδους. Οι κατηγορίες στην ταξινόμηση είναι προκαθορισμένες, ενώ στην ομαδοποίηση οι κατηγορίες δεν είναι γνωστές εκ των προτέρων. Επίσης, η κατηγοριοποίηση είναι εποπτευόμενη, ενώ η ομαδοποίηση είναι μη-εποπτευομένη.

2.3.2 Εξόρυξη γνώμης

Η πρόσφατη ανάπτυξη των μέσων κοινωνικής δικτύωσης οδήγησε σε μια πρωτόγνωρη ανταλλαγή απόψεων για τα πιο διαφορετικά θέματα. Άνθρωποι από διάφορα μέρη του κόσμου δημοσιεύουν τις απόψεις τους για προϊόντα, μοιράζονται αξιολογήσεις ξενοδοχείων, δίνουν ταξιδιωτικές συμβουλές, μοιράζονται απόψεις σχετικά με την ανατροφή των παιδιών, συζητούν πολιτικές απόψεις, και ούτω καθεξής. Αυτό το φαινόμενο έχει καταστήσει πρακτικά αδύνατο να παρακολουθούνται όλες αυτές οι απόψεις που κυκλοφορούν στο διαδίκτυο. Για να αντιμετωπίσουν αυτό το πρόβλημα, ορισμένες εφαρμογές βασίζονται σε αυτόματες τεχνικές ανάλυσης απόψεων. Με αυτόν τον τρόπο, μπορεί να δημιουργηθεί μια πιο εκφραστική σύνθεση κειμένου που να εξηγεί το συναίσθημα που εκφράζεται σε ένα κείμενο, δημιουργώντας έτσι πιο φυσικές αλληλεπιδράσεις με τον ανθρώπινο υπολογιστή. Επιπλέον, η δημιουργία χρονοδιαγραμμάτων αλλαγών συναισθήματος σε διαδικτυακά φόρουμ ή ειδήσεις επιτρέπει σε κάποιον να παρακολουθεί τις κορυφές θετικού ή αρνητικού συναισθήματος στην κοινή γνώμη. Επιπλέον, η διαχείριση πελατειακών σχέσεων έχει λάβει σημαντική προσοχή από πολλές εταιρείες. Τα φόρουμ πελατών παρακολουθούνται για εκφράσεις αρνητικών συναισθημάτων σχετικά με ένα προϊόν ή μια επωνυμία, επιτρέποντας έτσι στο τμήμα εξυπηρέτησης πελατών να παρέμβει

έγκαιρα. Τέλος, οι απόψεις που αντλούνται από κριτικές προϊόντων μπορούν να ενημερώσουν ένα άτομο για τα πλεονεκτήματα και τα μειονεκτήματα ενός προϊόντος που σκέφτεται να αγοράσει. Αυτές οι απόψεις επίσης μπορούν να βοηθήσουν τις εταιρείες να κατανοήσουν τα δυνατά και αδύνατα σημεία των προϊόντων τους. Άλλες εφαρμογές που επωφελούνται από την εξόρυξη απόψεων είναι η απάντηση σε ερωτήσεις, η περίληψη συνομιλιών και η σημασιολογική ανάλυση κειμένου [27].

Επομένως η εξόρυξη γνώμης (Opinion mining) [35, 37, 36] ορίζεται ως ένα σημαντικό είδος πληροφοριών που μεταφέρεται σε πολλούς τύπους γραπτού και προφορικού λόγου και αφορά την ψυχική ή συναισθηματική κατάσταση του συγγραφέα, του ομιλητή ή άλλης οντότητας που αναφέρεται στον λόγο. Τα άρθρα ειδήσεων, για παράδειγμα, συχνά αναφέρουν συναισθηματικές αντιδράσεις σε μια ιστορία, εκτός από τα γεγονότα. Τα συντακτικά άρθρα, οι κριτικές, τα ιστολόγια και οι πολιτικές ομιλίες μεταφέρουν τις απόψεις, τις πεποιθήσεις ή τις προθέσεις του συγγραφέα, του ομιλητή ή της συγκεκριμένης οντότητας. Ένας μαθητής που συμμετέχει σε μια συνεδρία διδασκαλίας μπορεί να εκφράσει την κατανόησή του ή την αβεβαιότητά του. Οι απόψεις είναι εκφράσεις προσωπικών καταστάσεων, όπως συναισθήματα, αξιολογήσεις, πεποιθήσεις και εικασίες σε φυσική γλώσσα. Οι απόψεις διαθέτουν χαρακτηριστικά όπως ποιος εκφράζει τη γνώμη, το είδος της στάσης που εκφράζεται, για ποιον ή τι εκφράζεται η γνώμη, το συναίσθημα της γνώμης (δηλαδή εάν είναι θετική ή αρνητική) και άλλα.

Η εξόρυξη γνώμης συνήθως χωρίζεται σε δύο κύριες επιμέρους εργασίες:

1. Ανάλυση υποκειμενικότητας, η οποία προσδιορίζει εάν ένα κείμενο περιέχει μια γνώμη και αντίστοιχα χαρακτηρίζει το κείμενο ως υποκειμενικό ή αντικειμενικό.

2. Ανάλυση συναισθήματος, η οποία ταξινομεί περαιτέρω μια γνώμη (ή υποκειμενικό κείμενο) ως θετικό, αρνητικό ή ουδέτερο.

Η κατανόηση της γνώμης που εκφράζεται σε δεδομένα κειμένου είναι απαραίτητη για τις επιχειρήσεις για τη μέτρηση της ικανοποίησης των πελατών, της κοινής γνώμης και της αντίληψης της επωνυμίας. Βοηθά στην ανάλυση των κριτικών, του περιεχομένου των μέσων κοινωνικής δικτύωσης και των σχολίων των πελατών για την εξαγωγή συναισθημάτων όπως θετικά, αρνητικά ή ουδέτερα.

2.3.3 Εξαγωγή Πληροφοριών

Συγκεκριμένες πληροφορίες, όπως ονόματα οντοτήτων ή οργανισμών, ή οι σχέσεις μεταξύ τους, συχνά θάβονται σε μη δομημένο κείμενο. Για παράδειγμα, στη φράση “Ο Στίβ Τζόμπς, ιδρυτής της Apple, παρουσίασε το νέο iPhone στην εκδήλωση της εταιρείας”, περιλαμβάνονται αναφορές σε πολλές οντότητες όπως ονόματα οργανισμών (Apple), το όνομα ενός ατόμου (Στίβ Τζόμπς), και μια εκδήλωση

(παρουσίαση του νέου iPhone). Επιπλέον, περιγράφεται ένα γεγονός που αφορά την παρουσίαση του νέου προϊόντος. Στο συγκεκριμένο κείμενο, εκτός από την αναφορά σε μια συγκεκριμένη εκδήλωση, παρουσιάζονται επίσης αναφορές σε σχέσεις μεταξύ οντοτήτων, όπως η σχέση μεταξύ του Στίβ Τζόμπς και της Apple.

Η εξαγωγή πληροφοριών (Information extraction) [27, 31, 35] είναι η διαδικασία εξαγωγής δομημένων πληροφοριών από μη δομημένα δεδομένα. Ο τύπος των πληροφοριών που εξάγονται είναι συνήθως προκαθορισμένος και περιλαμβάνει οντότητες όπως άτομα ή ονόματα επιχειρήσεων, εκδηλώσεις όπως παρουσιάσεις προϊόντων ή συμβάντα, και σχέσεις όπως η σύνδεση μεταξύ προσώπων και εταιρειών. Παράλληλα, πρόσφατα έχουν αναπτυχθεί συστήματα που αποσκοπούν στην “Ανοικτή Εξαγωγή Πληροφοριών”, όπου οι πληροφορίες που εξάγονται δεν είναι προκαθορισμένες. Τα συστήματα Εξαγωγής Πληροφοριών έχουν πολλές πρακτικές εφαρμογές. Μπορούν, για παράδειγμα, να χρησιμοποιηθούν για την εξαγωγή ονομάτων ασθενειών, συμπτωμάτων ή φαρμάκων, την αναγνώριση εταιρειών και των διευθυντών τους, των ονομάτων καθηγητών και των πανεπιστημίων στα οποία διδάσκουν, και άλλων παρόμοιων πληροφοριών. Μπορούν επίσης να χρησιμοποιηθούν για τη συλλογή δεδομένων σχετικά με τις σχέσεις μεταξύ οντοτήτων, τις τάσεις σε διάφορους τομείς, και άλλες παρόμοιες αναλύσεις.

2.3.4 Ανάλυση Θεμάτων

Οι συνομιλίες μπορούν να αναλυθούν με διάφορους τρόπους, συμπεριλαμβανομένων των αφηγήσεων, των θεμάτων και των μεταφορών [27]. Καθώς τα γεγονότα εξελίσσονται, διαφορετικά μέσα ενημέρωσης, με διαφορετικό κοινό και διαφορετικές επιχειρηματικές στρατηγικές, επιλέγουν να εστιάσουν σε διαφορετικά θέματα. Μέσα ενημέρωσης με μεγάλη επιρροή, καθώς και πρόσωπα με εξουσία, όπως πολιτικοί και διάσημοι, διαδραματίζουν σημαντικό ρόλο στη διαμόρφωση των θεμάτων που εμφανίζονται στα μέσα ενημέρωσης [46]. Τα θέματα που καλύπτονται από τα μέσα με την σειρά τους επηρεάζουν τις στάσεις και τις απόψεις των θεατών και των αναγνωστών. Οι συμπεριφορές και οι απόψεις του κοινού θα επηρεάσουν τις προτιμήσεις τους στις εκλογές και τις αποφάσεις που θα λαμβάνονται από εκλεγμένους αξιωματούχους. Ένα άλλο παράδειγμα είναι η σημασία του θέματος που θα επιλέξει κάποιος σε μια συνομιλία σε διαφορετικά περιβάλλοντα. Φανταστείτε ότι μετά την αποφοίτησή ενός φοιτητή από το πανεπιστήμιο, επρόκειτο να δεχτεί μια προσφορά εργασίας από μια εταιρεία μάρκετινγκ. Για να πετύχει κάποιος στην νέα δουλειά, δεν θα πρέπει μόνο να μάθει τις τεχνικές απαιτήσεις της θέσης, αλλά και να αναπτύξει καλές σχέσεις με τους νέους συναδέλφους του. Άρα για να προσαρμοστεί κανείς σε μια νέα δουλειά θα πρέπει να αλλάξει τα θέματα που συνήθως συζητάει. Ενώ ένας φοιτητής συνήθως συζητάει θέματα όπως μουσική, ταινίες, μαθήματα, εργασίες, καθηγητές, σχέσεις και πολιτική, στη νέα δουλειά, αυτά τα θέματα αλλάζουν. Με τη

μεταβολή του περιβάλλοντος εργασίας, αλλάζουν και οι συζητήσεις. Αυτό απαιτεί να μιλήσετε για θέματα όπως έργα εργασίας, καριέρες, πελάτες και κάθε άλλο θέμα που είναι σχετικό με το νέο περιβάλλον. Η ικανότητα κάποιου να ανταποκρίνεται στα θέματα που συζητούνται στο εργασιακό περιβάλλον συνδέεται άμεσα με τις προηγούμενες εμπειρίες του στη ζωή. Αυτή η δεξιότητα αποτελεί αντικείμενο μελέτης για κοινωνικούς επιστήμονες που ερευνούν τις σχέσεις μεταξύ σταδιοδρομίας, επαγγελμάτων και κοινωνικών ανισοτήτων.

Για να εντοπίσουμε ποια θέματα συζητούνται εντός μιας κοινότητας ή μιας κοινωνικής ομάδας, δεν χρειάζεται πλέον να διαβάσουμε χιλιάδες άρθρα εφημερίδων ή να ακούσουμε αμέτρητες συνεντεύξεις. Η ανάγνωση τόσων μεγάλων όγκων κειμένων είναι χρονοβόρα και, επειδή είναι επιρρεπής σε υποκειμενικές ερμηνείες, δεν θεωρείται επιστημονικά έγκυρη από μόνη της. Αντί γι' αυτό, οι ερευνητές έχουν αναπτύξει τα τελευταία χρόνια μοντέλα θεμάτων (Analyzing topics) [41, 42, 45, 34]. Αυτά τα μοντέλα είναι στατιστικά, και χρησιμοποιούνται για να ανιχνεύσουν τους συνδυασμούς των θεμάτων που συζητούνται σε μια κοινωνική ομάδα. Επίσης, βοηθούν στο να κατανοήσουμε πώς μεταβάλλονται τα θέματα που συζητούνται με την πάροδο του χρόνου. Οι αλγόριθμοι μοντελοποίησης θεμάτων [43, 44, 47] ανακαλύπτουν αυτόματα αφηρημένα θέματα από μια συλλογή εγγράφων. Βοηθά στην οργάνωση, τη σύνοψη και την κατανόηση μεγάλων σωμάτων εντοπίζοντας υποκείμενα θέματα ή θέματα στα έγγραφα.

2.3.5 Σύνοψη Κειμένου

Πριν πάμε στη σύνοψη κειμένου (Text Summarization) [33, 35, 38], πρώτα πρέπει να γνωρίζεται τι είναι η περίληψη. Η περίληψη αναφέρει τις κύριες πληροφορίες ενός ή περισσότερων κειμένων σε συνοπτική μορφή, παρέχοντας μια συνοπτική επισκόπηση του αρχικού κειμένου. Ο στόχος της σύνοψης κειμένου είναι η παρουσίαση του κειμένου πηγής σε μια συντομότερη έκδοση με σημασιολογία η με λίγα λογία η σύνοψη βοηθά στη γρήγορη αφομοίωση τεράστιων ποσοτήτων πληροφοριών, διευκολύνοντας τη λήψη αποφάσεων και την ανάκτηση πληροφοριών. Το κυριότερο πλεονέκτημα της χρήσης περίληψης είναι η εξοικονόμηση χρόνου ανάγνωσης. Οι μέθοδοι σύνοψης κειμένου χωρίζονται σε δύο κατηγορίες: εξαγωγική και αφηρημένη.

- Η εξαγωγική μέθοδος σύνοψης περιλαμβάνει την επιλογή σημαντικών προτάσεων, παραγράφων κλπ. από το αρχικό έγγραφο και τη σύνταξή τους σε συντομότερη μορφή.
- Η αφηρημένη σύνοψη αναφέρεται στην κατανόηση των κύριων εννοιών ενός κειμένου και στη συνέχεια στην έκφρασή τους με σαφή φυσική γλώσσα.

Οι δύο κύριες κατηγορίες σύνοψης κειμένου είναι η ενδεικτική και η ενημερωτική. Η ενδεικτική περίληψη αντιπροσωπεύει μόνο την κύρια ιδέα του κειμένου. Το τυπικό μήκος αυτής της περίληψης είναι 5 έως 10% του αρχικού κειμένου. Από την άλλη πλευρά, οι ενημερωτικές περιλήψεις παρέχουν συνοπτικές πληροφορίες για το κύριο κείμενο. Το μήκος της ενημερωτικής περίληψης είναι 20 έως 30% του ίδιου κειμένου. Υπάρχουν τρία βασικά βήματα για τη σύνοψη των κειμένων. Αυτά είναι ο προσδιορισμός θέματος, η ερμηνεία και η δημιουργία περιλήψεων.

- Κατά τον προσδιορισμό του θέματος εντοπίζονται οι πιο σημαντικές πληροφορίες στο κείμενο. Χρησιμοποιούνται διάφορες τεχνικές για τον προσδιορισμό του θέματος, οι οποίες είναι η θέση, οι φράσεις υπόδειξης, η συχνότητα λέξεων. Οι μέθοδοι που βασίζονται στη θέση των φράσεων αποτελούν τις πιο αποτελεσματικές προσεγγίσεις για τον προσδιορισμό του θέματος ενός κειμένου.

- Στην ερμηνεία οι αφηρημένες περιλήψεις πρέπει να περάσουν από το στάδιο της ερμηνείας. Σε αυτό το βήμα, διάφορα θέματα συγχωνεύονται για να σχηματίσουν ένα γενικό περιεχόμενο.

- Στην δημιουργία περιλήψεων το σύστημα χρησιμοποιεί τη μέθοδο δημιουργίας κειμένου.

Στο παρελθόν, οι εξαγωγικοί συνοψιστές βασίζονταν κυρίως στην αξιολόγηση των προτάσεων στο αρχικό έγγραφο. Οι πιο σύγχρονες τεχνικές σύνοψης κειμένου χρησιμοποιούν είτε στατιστικές προσεγγίσεις είτε γλωσσικές τεχνικές. Λέξεις με υψηλή συχνότητα, η μέθοδος υπόδειξης, η τυπική λέξη-κλειδί, η μέθοδος τίτλου και η μέθοδος θέσης χρησιμοποιούνται για τον προσδιορισμό της σημασίας των προτάσεων.

2.4 Μοντέλα

Τα μοντέλα που χρησιμοποιούνται στην εξόρυξη κειμένου είναι ποικίλα και συχνά εξειδικευμένα για συγκεκριμένες εφαρμογές. Ορισμένα από τα κύρια μοντέλα περιλαμβάνουν:

2.4.1 Νευρωνικά Δίκτυα

Η έμπνευση για τα νευρωνικά δίκτυα (Neural Networks) [15, 31, 36] ήταν η αναγνώριση ότι τα πολύπλοκα συστήματα μάθησης στον εγκέφαλο των ζώων αποτελούνταν από στενά διασυνδεδεμένα σύνολα νευρώνων. Παρόλο που ένας μεμονωμένος νευρώνας είναι απλής δομής, τα πυκνά δίκτυα συνδεδεμένων νευρώνων μπορούν να αναλάβουν πολύπλοκες εργασίες μάθησης, όπως η ταξινόμηση και η αναγνώριση προτύπων [27]. Για παράδειγμα, ο ανθρώπινος εγκέφαλος περιέχει περίπου 10^11 νευρώνες, και κάθε

νευρώνας συνδέεται κατά μέσο όρο με 10.000 άλλους νευρώνες, κάνοντας συνολικά περίπου 10^15 συνδέσεις. Τα τεχνητά νευρωνικά δίκτυα (εφεξής, νευρωνικά δίκτυα) προσπαθούν, σε πολύ βασικό επίπεδο, να αντιγράψουν τον τρόπο μη γραμμικής μάθησης που παρατηρείται στα δίκτυα νευρώνων στη φύση [35]. Ανήκουν σε μια άλλη κατηγορία μη γραμμικών μεθόδων πρόβλεψης και ταξινόμησης [33, 35]. Τα νευρωνικά δίκτυα λειτουργούν υπολογιστικά, μιμούμενα τη λειτουργία των νευρώνων. Ένα τυπικό νευρωνικό δίκτυο αποτελείται από στρώματα εισόδου και εξόδου, τα οποία περιλαμβάνουν πολλές μονάδες, με τις συνδέσεις μεταξύ των μονάδων να έχουν βάρη που σχετίζονται με αυτές. Συγκεκριμένα, ένα πολυεπίπεδο νευρωνικό δίκτυο αποτελείται από:

- Ένα στρώμα εισόδου με διάφορες μονάδες εισόδου. Συνήθως, αυτές οι μονάδες εισόδου αντιστοιχούν στα χαρακτηριστικά της περίπτωσης που πρέπει να ταξινομηθεί.
- Ένα ή περισσότερα κρυφά επίπεδα που αποτελούνται από πολλές μονάδες που μοιάζουν με νευρώνες.
- Ένα στρώμα εξόδου με πολλές μονάδες, οι οποίες αντιπροσωπεύουν τις κλάσεις που πρέπει να προβλεφθούν.

Κάθε μονάδα λαμβάνει ως είσοδο έναν συνδυασμένο γινόμενο των εισόδων του προηγούμενου στρώματος, το οποίο ζυγίζεται με βάρη. Έπειτα, εφαρμόζεται μια μη γραμμική συνάρτηση ενεργοποίησης για τον υπολογισμό της εξόδου της μονάδας. Για συνάρτηση ενεργοποίησης συνήθως χρησιμοποιείται μια λογιστική συνάρτηση, η οποία μετατρέπει την έξοδο σε έναν αριθμό μεταξύ 0 και 1 επομένως, συχνά αναφέρεται ως συνάρτηση σιγμοειδούς. Οι μονάδες εξόδου μπορεί να αντιπροσωπεύουν είτε την πραγματική έξοδο του συστήματος, είτε τις εξόδους του πρώτου κρυφού στρώματος οι οποίες μπορούν να τροφοδοτηθούν ως είσοδοι σε ένα άλλο κρυφό στρώμα εάν χρειαστεί να εμπλακούν περισσότερα κρυφά στρώματα.

Χρειάζεται να αποφασιστούν τα βάρη που συνδέουν τις εισόδους και τις εξόδους αυτών των στρωμάτων, ο αριθμός των μονάδων στα διάφορα επίπεδα και ο αριθμός των κρυφών επίπεδων. Τα νευρωνικά δίκτυα είναι πολύ γενικά, καθώς η χρήση μη γραμμικών συναρτήσεων ενεργοποίησης τα καθιστά ικανά να προσεγγίσουν με ακρίβεια πολλές μη γραμμικές συναρτησιακές σχέσεις. Φυσικά, τα βάρη σε αυτά τα συστήματα πρέπει να προσαρμοστούν κατά τη διάρκεια της εκπαίδευσης στα πραγματικά δεδομένα. Η μέθοδος της αντίστροφης διάδοσης μπορεί να εφαρμοστεί επαναληπτικά για την προσαρμογή των βαρών με σκοπό τη βελτίωση της απόδοσης σε ένα δεδομένο σύνολο εκπαίδευσης. Γι' αυτό, είναι απαραίτητο να οριστεί ένας παράγοντα μάθησης που ρυθμίζει την ταχύτητα σύγκλισης της επαναληπτικής μεθόδου εκτίμησης.

Τα νευρωνικά δίκτυα είναι πολύ ευέλικτα και μπορούν να προσεγγίσουν περίπλοκες σχέσεις. Ένα από τα σημαντικά πλεονεκτήματα [6, 15] της χρήσης νευρωνικών δικτύων είναι η ικανότητά τους να αντιμετωπίζουν αποτελεσματικά θορυβώδη δεδομένα. Εξαιτίας του μεγάλου αριθμού κόμβων (τεχνητών νευρώνων) και των βαρών που αντιστοιχούν σε κάθε σύνδεση, το δίκτυο μπορεί να μάθει να αντιμετωπίζει ακόμα και τα μη ενδεικτικά (ή ακόμα και εσφαλμένα) παραδείγματα που περιέχονται στο σύνολο δεδομένων. Ωστόσο, έχουν και ορισμένα μειονεκτήματα [6, 15]. Ένα από αυτά είναι ότι τα προσεγγιστικά μοντέλα που παράγονται από νευρωνικά δίκτυα είναι αδιαφανή (“μαύρα κουτιά”) και παρέχουν περιορισμένη κατανόηση σχετικά με τον τρόπο λειτουργίας τους. Επιπλέον, ο χρήστης νευρωνικών δικτύων πρέπει να κάνει πολλές υποθέσεις σχετικά με τη μοντελοποίηση τους. Αυτές οι υποθέσεις συμπεριλαμβάνουν τον αριθμό των κρυφών επιπέδων, τον αριθμό των μονάδων σε κάθε επίπεδο, τον τύπο των συναρτήσεων ενεργοποίησης και τους τρόπους εκπαίδευσης και συνήθως λείπει κατευθυντήρια γραμμή για τη λήψη τέτοιων αποφάσεων. Η εύρεση της καταλληλότερης αναπαράστασης απαιτεί σημαντική εμπειρία. Επιπλέον, τα νευρωνικά δίκτυα συνήθως απαιτούν μεγάλους χρόνους εκπαίδευσης που συχνά εκτείνονται σε αρκετές ώρες με συνέπεια να είναι αρκετά. αργή εάν η σταθερά μάθησης δεν επιλεγεί σωστά.

2.4.2 Bag-of-Words

Το μοντέλο Bag-of-Words (BoW) [31, 32, 40] είναι μια θεμελιώδης τεχνική που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας (NLP) και στην ανάκτηση πληροφοριών. Αντιπροσωπεύει τα δεδομένα ενός κειμένου μετρώντας τη συχνότητα εμφάνισης των λέξεων που περιέχει το έγγραφο. Αγνοεί τη γραμματική, τη σειρά των λέξεων και τη δομή, αντιμετωπίζοντας κάθε έγγραφο ως ένα “σακούλι” λέξεων. Έτσι, απλοποιεί πολύπλοκα έγγραφα σε ένα μοντέλο διανυσματικού χώρου. Η πρώτη διαδικασία στη μέθοδο BoW είναι η δημιουργία ενός λεξιλογίου, το οποίο περιλαμβάνει όλες τις μοναδικές λέξεις που υπάρχουν στο σώμα κειμένων (συλλογή εγγράφων). Κάθε λέξη αντιστοιχίζεται σε ένα μοναδικό ευρετήριο ή αναγνωριστικό. Μόλις δημιουργηθεί το λεξιλόγιο, κάθε έγγραφο αναπαρίσταται ως ένα διάνυσμα, όπου κάθε στοιχείο του διανύσματος αντιστοιχεί στη συχνότητα μιας λέξης από το λεξιλόγιο που υπάρχει στο έγγραφο. Αυτή η διαδικασία είναι επίσης γνωστή ως εξαγωγή χαρακτηριστικών. Τα διανύσματα BoW είναι συνήθως αραιά επειδή τα περισσότερα έγγραφα περιέχουν μόνο ένα μικρό υποσύνολο ολόκληρου του λεξιλογίου. Αυτή η αραιότητα μπορεί να οδηγήσει σε υπολογιστικές προκλήσεις και αναποτελεσματικότητα, ειδικά με μεγάλα λεξιλόγια. Το BoW μετατρέπει έγγραφα κειμένου σε διανύσματα υψηλών διαστάσεων, όπου κάθε διάσταση αντιπροσωπεύει μια μοναδική λέξη στο λεξιλόγιο. Κατά συνέπεια, τα έγγραφα μπορούν να συγκριθούν με βάση τις διανυσματικές αναπαραστάσεις τους χρησιμοποιώντας μέτρα ομοιότητας όπως η ομοιότητα συνημίτονου. Χρησιμοποιείται συνήθως σε εργασίες όπως η

ταξινόμηση και η ομαδοποίηση εγγράφων. Στην ταξινόμηση, οι αλγόριθμοι μηχανικής μάθησης εκπαιδεύονται σε αναπαραστάσεις BoW εγγράφων με ετικέτα για να προβλέψουν τις ετικέτες νέων η χωρίς ετικέτα εγγράφων. Στην ομαδοποίηση, έγγραφα με παρόμοιες αναπαραστάσεις BoW ομαδοποιούνται.

Τα βήματα προ-επεξεργασίας [29], όπως ο συμβολισμός (tokenization), η διακοπή της αφαίρεσης λέξης (the termination of word subtraction), η αποκοπή (stemming) και η λημματοποίηση (lemmatization) εφαρμόζονται συχνά στα δεδομένα κειμένου πριν από τη δημιουργία της αναπαράστασης BoW. Αυτά τα βήματα συμβάλλουν στη βελτίωση της ποιότητας της αναπαράστασης και στη μείωση της διάστασης του χώρου χαρακτηριστικών. Παρά την απλότητα και την αποτελεσματικότητά του σε ορισμένες εργασίες, το μοντέλο BoW έχει περιορισμούς. Δεν καταγράφει σημασιολογικές σχέσεις μεταξύ λέξεων, αγνοεί τη σειρά των λέξεων και αντιμετωπίζει όλες τις λέξεις εξίσου ανεξάρτητα από τη σημασία ή τη σημασία τους στο έγγραφο. Διάφορες επεκτάσεις και βελτιώσεις στο βασικό μοντέλο BoW έχουν προταθεί για την αντιμετώπιση των περιορισμών του. Αυτές περιλαμβάνουν τεχνικές όπως TF-IDF (Term Frequency-Inverse Document Frequency), ενσωματώσεις λέξεων (π.χ. Word2Vec, GloVe) και μοντέλα που βασίζονται σε νευρωνικά δίκτυα όπως μετασχηματιστές (π.χ. BERT), τα οποία καταγράφουν πιο λεπτές σημασιολογικές πληροφορίες [39]. Συνοπτικά, το μοντέλο BoW είναι μια θεμελιώδης τεχνική στο NLP, παρέχοντας έναν απλό αλλά ισχυρό τρόπο αναπαράστασης και ανάλυσης δεδομένων κειμένου.

2.4.3 Ενσωματώσεις Λέξεων

Το μοντέλο ενσωματώσεων λέξεων (Word Embeddings) [22, 27, 48] είναι μια μετασχηματιστική τεχνική στην επεξεργασία φυσικής γλώσσας (NLP) [49, 50] που αναπαριστά τις λέξεις ως πυκνά διανύσματα σε έναν συνεχή διανυσματικό χώρο. Σε αντίθεση με το μοντέλο Bag-of-Words, το οποίο αναπαριστά τις λέξεις ως αραιά διανύσματα συχνοτήτων όρων, οι ενσωματώσεις λέξεων αντιπροσωπεύουν τις λέξεις ως πυκνά διανύσματα συνεχών τιμών. Κάθε λέξη αντιστοιχίζεται σε ένα διάνυσμα σταθερού μεγέθους, συνήθως με εκατοντάδες διαστάσεις. Ένα από τα πιο σημαντικά πλεονεκτήματα των ενσωματώσεων λέξεων είναι η ικανότητά τους να καταγράφουν σημασιολογικές σχέσεις μεταξύ των λέξεων. Οι λέξεις με παρόμοια σημασία αντιπροσωπεύονται από διανύσματα που βρίσκονται κοντά μεταξύ τους στον χώρο ενσωμάτωσης, επιτρέποντας πιο λεπτή κατανόηση της σημασιολογίας των λέξεων. Βασίζονται στην υπόθεση διανομής, η οποία υποστηρίζει ότι οι λέξεις που εμφανίζονται σε παρόμοια συμφραζόμενα τείνουν να έχουν παρόμοια σημασία και μαθαίνονται εκπαιδεύοντας τα νευρωνικά δίκτυα σε μεγάλα σώματα κειμένου για να προβλέψουν το περιεχόμενο κάθε λέξης με βάση τις λέξεις που την περιβάλλουν. Οι προ-εκπαιδευμένες ενσωματώσεις λέξεων χρησιμοποιούνται ευρέως σε εργασίες NLP. Μοντέλα όπως το Word2Vec, το GloVe και το FastText έχουν εκπαιδευτεί σε μεγάλα

σώματα κειμένου και είναι διαθέσιμα για χρήση σε διάφορες εφαρμογές. Αυτές οι προ εκπαιδευμένες ενσωματώσεις καταγράφουν γενικές σημασιολογικές σχέσεις και μπορούν να βελτιστοποιηθούν σε συγκεκριμένες εργασίες ή τομείς.

Οι ενσωματώσεις λέξεων έχουν συνήθως πολύ χαμηλότερη διάσταση σε σύγκριση με το μοντέλο Bag-of-Words, καθιστώντας τις πιο αποδοτικές υπολογιστικά και ευκολότερες στην εργασία. Παρά το γεγονός ότι έχουν λιγότερες διαστάσεις, οι ενσωματώσεις λέξεων μπορούν ακόμα να συλλάβουν πολύπλοκες σημασιολογικές πληροφορίες. Η ομοιότητα μεταξύ λέξεων στον χώρο ενσωμάτωσης μετριέται συχνά χρησιμοποιώντας ομοιότητα συνημίτονου ή ευκλείδεια απόσταση. Αυτό επιτρέπει εργασίες όπως υπολογισμός ομοιότητας λέξεων, ανίχνευση αναλογίας (π.χ. “βασιλιάς” - “άνδρας” + “γυναίκα” ≈ “βασίλισσα”) και ομαδοποίηση λέξεων με βάση τη σημασιολογική ομοιότητα.

Οι ενσωματώσεις λέξεων που έχουν εκπαιδευτεί σε μεγάλα σώματα κειμένου μπορούν να χρησιμοποιηθούν ως αναπαραστάσεις χαρακτηριστικών για μεταγενέστερες εργασίες NLP, όπως η ανάλυση συναισθήματος, η αναγνώριση ονομαστικών οντοτήτων και η αυτόματη μετάφραση. Αυτή η προσέγγιση εκμάθησης μεταφοράς αξιοποιεί τις σημασιολογικές πληροφορίες που συλλαμβάνονται στις ενσωματώσεις για να βελτιώσει την απόδοση των μοντέλων για συγκεκριμένες εργασίες, ειδικά όταν τα δεδομένα με ετικέτα είναι περιορισμένα.

Οι πρόσφατες εξελίξεις στις ενσωματώσεις λέξεων περιλαμβάνουν ενσωματώσεις με βάση τα συμφραζόμενα, όπως ELMo, GPT και BERT. Αυτά τα μοντέλα συλλαμβάνουν έννοιες λέξεων που εξαρτώνται από το πλαίσιο λαμβάνοντας υπόψη ολόκληρη την πρόταση ή το έγγραφο, με αποτέλεσμα πιο ακριβείς αναπαραστάσεις, ειδικά για εργασίες που απαιτούν κατανόηση πολύπλοκων γλωσσικών δομών. Ενώ οι ενσωματώσεις λέξεων είναι ισχυρά εργαλεία για τη λήψη σημασιολογικών πληροφοριών, δεν είναι χωρίς περιορισμούς. Μπορεί να δυσκολεύονται με σπάνιες λέξεις ή λέξεις εκτός λεξιλογίου και μπορεί επίσης να κωδικοποιούν προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης. Επιπλέον, το πλαίσιο που καταγράφεται από τις ενσωματώσεις λέξεων περιορίζεται στο τοπικό πλαίσιο στο οποίο εμφανίζονται οι λέξεις στα δεδομένα εκπαίδευσης. Συνοπτικά, οι ενσωματώσεις λέξεων έχουν φέρει επανάσταση στον τομέα του NLP παρέχοντας πυκνές, σημασιολογικά σημαντικές αναπαραστάσεις λέξεων που επιτρέπουν πιο εξελιγμένη κατανόηση και επεξεργασία της γλώσσας. Η ευελιξία και η αποτελεσματικότητά τους τα καθιστούν ακρογωνιαίο λίθο της σύγχρονης έρευνας και εφαρμογών NLP όπως είναι η αυτόματη μετάφραση μεταξύ διαφορετικών γλωσσών, διευκολύνοντας την επικοινωνία και την ανταλλαγή πληροφοριών μεταξύ διαφορετικών γλωσσικών κοινοτήτων και η δημιουργία κειμένου που μοιάζει με άνθρωπο, συμπεριλαμβανομένων των απαντήσεων chatbot, της δημιουργίας περιεχομένου και της δημιουργικής γραφής, επιτρέποντας διάφορες εφαρμογές συνομιλίας και δημιουργικότητας.

2.4.4 Μηχανική και Βαθιά Μάθηση

Τα μοντέλα μηχανικής μάθησης (Machine Learning) [51-54], όπως το Naive Bayes, Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) και οι Κ-Πλησιέστεροι Γείτονες (K-Nearest Neighbors - KNN), ανήκουν στην κατηγορία της παραδοσιακής μηχανικής μάθησης. Αυτά τα μοντέλα είναι σχετικά εύκολα στην υλοποίηση και κατανόηση, και συνήθως λειτουργούν καλά για απλά προβλήματα εξόρυξης κειμένου. Για παράδειγμα, το Naive Bayes υποθέτουν ανεξαρτήτως της πραγματικότητας ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, οι SVM επιδιώκουν να βρουν τον καλύτερο διαχωρισμό μεταξύ δύο κατηγοριών, ενώ οι KNN βασίζονται στην απόσταση μεταξύ δειγμάτων για να προβλέψουν την κατηγορία ενός νέου δείγματος.

Από την άλλη πλευρά, τα μοντέλα βαθιάς μάθησης (Deep Learning) [27, 28, 33, 52] είναι πολύ πιο πολύπλοκα και περιλαμβάνουν πολλά επίπεδα νευρωνικών δικτύων. Αυτά τα μοντέλα εκμεταλλεύονται τη δύναμη των υπολογιστικών πόρων και των μεγάλων συνόλων δεδομένων για να εξάγουν πιο πολύπλοκες και αφηρημένες αναπαραστάσεις από τα δεδομένα εισόδου. Αυτά τα μοντέλα είναι κατάλληλα για εργασίες όπως η ανάλυση συναισθημάτων, η αυτόματη κατανόηση φυσικής γλώσσας και άλλα πολύπλοκα προβλήματα επεξεργασίας γλώσσας και αναγνώρισης προτύπων.

3. Αλγόριθμοι Εξαγωγής Λέξεων/Φράσεων Κλειδιών

Οι αλγόριθμοι εξόρυξης κειμένου (Text Mining Algorithms) [55] είναι αλγόριθμοι που χρησιμοποιούνται για την ανάλυση και την εξαγωγή πληροφορίας από κείμενο [57]. Η εξόρυξη κειμένου είναι ένας τομέας της τεχνητής νοημοσύνης που εστιάζει στην ανάλυση μεγάλων όγκων κειμένου για την ανακάλυψη μοτίβων, τάσεων, πληροφορίας και γνώσης. Χρησιμοποιούνται σε κοινωνικά δίκτυα, κοινωνικής δικτύωσης, επιχειρήσεις, εμπορία, δημοσιογραφία, ΜΜΕ, υγειά, ακαδημαϊκή ερευνά, πολιτική και κυβερνητικές εφαρμογές. Οι λόγοι που χρησιμοποιούνται είναι για την ανάλυση συναισθημάτων, την αναγνώριση κλειδιών θεμάτων, τον εντοπισμό τάσεων στα μηνύματα που δημοσιεύονται στα κοινωνικά δίκτυα, ανάλυση αναφορών πελατών, των αναθεωρήσεων προϊόντων και των συζητήσεων στα φόρουμ προκειμένου να ληφθούν αποφάσεις σχετικά με την πολιτική προϊόντων, δράσεις μάρκετινγκ, παρακολούθηση τάσεων ειδήσεων, ανάλυση συναισθήματος του κοινού, δημιουργία αναφορών, ανάλυση κλινικών αναφορών και των σχόλιων των ασθενών για την παρακολούθηση της δημοσίας υγείας, ανακάλυψη τάσεων σε ασθένειες, ανάλυση βιβλιογραφίας, εξαγωγή γνώσης από επιστημονικά άρθρα, ανακάλυψη νέων πληροφοριών, παρακολούθηση της δημοσίας γνώμης, εντοπισμό επικίνδυνων καταστάσεων η την ανίχνευση πιθανών απειλών. Αυτά είναι μερικά παραδείγματα των πολλών τομέων όπου οι αλγόριθμοι εξόρυξης κειμένου μπορούν να εφαρμοστούν για την ανάλυση και την εξαγωγή πληροφοριών από κείμενο. Υπάρχουν διάφοροι αλγόριθμοι εξόρυξης σημαντικών λέξεων-φράσεων οι οποίοι είναι:

3.1 Yet Another Keyword Extractor

Ο αλγόριθμος Yet Another Keyword Extractor (YAKE) [58] είναι ένας αλγόριθμος εξαγωγής λέξεων-κλειδιών μη-επιβλεπόμενης μάθησης, ανεξάρτητος από το περιεχόμενο, τον τομέα και τη γλώσσα των κειμένων. Βασίζεται στη σημασιολογική τους σημασία και στη συχνότητα εμφάνισής τους. Είναι γνωστός για την απλότητα και την αποτελεσματικότητά του, καθιστώντας τον κατάλληλο για διάφορες εργασίες επεξεργασίας φυσικής γλώσσας. Λαμβάνει ένα κείμενο ως είσοδο, καθώς και τις ακόλουθες παραμέτρους: μια τιμή για το μέγεθος του παραθύρου w (χρησιμοποιούμενη από ένα από τα στατιστικά χαρακτηριστικά), τον αριθμό των n-γραμμάτων (n-gram), το κατώφλι αποδιπλασιασμού και τη γλώσσα του κειμένου (η παράμετρος γλώσσας του κειμένου είναι απαραίτητη μόνο για τον καθορισμό της αντίστοιχης λίστας λέξεων). Ο αλγόριθμος YAKE αποτελείται από πέντε βασικά βήματα:

1. Προεπεξεργασία κειμένου και αναγνώριση υποψηφίων όρων: Αρχικά, το κείμενο διαιρείται σε προτάσεις και στη συνέχεια σε κομμάτια. Αφαιρούνται θορυβώδεις πληροφορίες όπως σημεία στίξης και διευθύνσεις URL, και κάθε διακριτικό(token) μετατρέπεται σε πεζό και σχολιάζεται με οριοθετείς ετικέτες. Το αποτέλεσμα είναι

ένας κατάλογος προτάσεων, όπου κάθε πρόταση χωρίζεται σε κομμάτια που περιέχουν σχολιασμένους όρους.

2. Εξαγωγή χαρακτηριστικών: Μετά την προεπεξεργασία κειμένου και την αναγνώριση υποψηφίων όρων, πραγματοποιείται στατιστική ανάλυση του κειμένου. Υπολογίζονται η συχνότητα εμφάνισης των όρων, το ευρετήριο προτάσεων όπου εμφανίζονται οι όροι, η συχνότητα εμφάνισης των ακρωνύμων και η συχνότητα εμφάνισης των κεφαλαίων όρων.
3. Υπολογισμός βαθμολογίας όρων: Υπολογίζεται η βαθμολογία του κάθε όρου με βάση τα στατιστικά στοιχεία που έχουν υπολογιστεί στο προηγούμενο βήμα.
4. Δημιουργία n-gram: Δημιουργείται μια λίστα υποψηφίων λέξεων-κλειδιών με βάση ένα συρόμενο παράθυρο μεγέθους n. Οι υποψήφιες λέξεις-κλειδιά σχηματίζονται με συνεχόμενες ακολουθίες όρων που ανήκουν στο ίδιο κομμάτι του κειμένου.
5. Βαθμολόγηση και εξαγωγή υποψήφιων λέξεων-κλειδιών: Τελικά, υπολογίζονται οι βαθμολογίες των υποψήφιων λέξεων-κλειδιών συνδυάζοντας τα χαρακτηριστικά κάθε λέξης. Η βαθμολογία υποδεικνύει πόσο πιθανό είναι μια λέξη να είναι λέξη-κλειδί. Με βάση τις βαθμολογίες που υπολογίστηκαν, ο YAKE επιλέγει τις λέξεις με τις υψηλότερες βαθμολογίες ως λέξεις-κλειδιά. Αυτές οι λέξεις αντιπροσωπεύουν τα σημαντικά θέματα ή έννοιες του κειμένου.

Με αυτόν τον τρόπο, ο αλγόριθμος YAKE εξάγει αποτελεσματικά λέξεις-κλειδιά από κείμενα, βασιζόμενος στη σημασιολογική τους σημασία και στη συχνότητα εμφάνισής τους.

3.1.1 Υλοποίηση στην Python

Η υλοποίηση στην python έχει γίνει με την βιβλιοθήκη ‘yake’ η οποία περιέχει έτοιμη την υλοποίηση του αλγορίθμου YAKE που περιγράφεται παραπάνω και εμείς απλώς την χρησιμοποιούμε. Ο τρόπος με τον οποίο το χρησιμοποιούμε φαίνεται στην παρακάτω εικόνα.

```
language = "en"
max_ngram_size = max_length
deduplication_threshold = 0.9
deduplication_algo = 'seqm'
windowSize = 1
numOfKeywords = 10000

custom_kw_extractor = yake.KeywordExtractor(lan=language, n=max_ngram_size,
  dedupLim=deduplication_threshold, dedupFunc=deduplication_algo, windowsSize=windowSize,
  top=numOfKeywords, features=None)

keywords = custom_kw_extractor.extract_keywords(extracted_text)
```

Αυτός ο κώδικας χρησιμοποιεί τη βιβλιοθήκη YAKE για τη δημιουργία ενός αντικειμένου KeywordExtractor με προσαρμοσμένες παραμέτρους. Έπειτα, εκτελεί τη μέθοδο extract_keywords() για την εξαγωγή των λέξεων-κλειδιών από το κείμενο extracted_text.

Εδώ είναι μια επεξήγηση για τις παραμέτρους που χρησιμοποιούνται:

- lan: Η γλώσσα του κειμένου.
- n: Το μέγιστο μέγεθος των n-γραμμάτων (η παράμετρος max_ngram_size στον κώδικα).
- dedupLim: Το κατώφλι αποδιπλασιασμού (η παράμετρος deduplication_threshold στον κώδικα).
- dedupFunc: Ο αλγόριθμος αποδιπλασιασμού (η παράμετρος deduplication_algo στον κώδικα).
- windowsSize: Το μέγεθος του παραθύρου.
- top: Ο αριθμός των λέξεων-κλειδιών που επιστρέφονται (η παράμετρος numOfKeywords στον κώδικα).
- features: Προαιρετικά χαρακτηριστικά που μπορούν να χρησιμοποιηθούν για την εξαγωγή λέξεων-κλειδιών.

Αυτός ο κώδικας θα επιστρέψει μια λίστα με τις λέξεις-κλειδιά που εξάγονται από το extracted_text.

3.2 TextRank

Το TextRank [59] είναι ένας αλγόριθμος κατάταξης βασισμένος σε γράφημα για εξαγωγή λέξεων-κλειδιών και προτάσεων στην επεξεργασία φυσικής γλώσσας. Εισήχθη από τους Mihalcea και Tarau το 2004 ως μέθοδος αυτόματης σύνοψης και εξαγωγής φράσεων. Ο αλγόριθμος είναι εμπνευσμένος από τον αλγόριθμο PageRank της Google, ο οποίος ταξινομεί τις ιστοσελίδες με βάση τη σημασία τους στον ιστό. Αυτός ο αλγόριθμος είναι συχνά χρήσιμος για την αυτόματη εξαγωγή λέξεων-κλειδιών ή σημαντικών φράσεων από μεγάλα κείμενα, χρησιμοποιώντας πληροφορίες σχετικά με τη δομή και το περιεχόμενο του κειμένου για να αποφασίσει τη σημασία των λέξεων ή των φράσεων. Παρακάτω ακολουθεί μια αναλυτική επεξήγηση του πώς λειτουργεί.

Πρώτον, το κείμενο υποβάλλεται σε διάφορες διαδικασίες προεπεξεργασίας, όπως κανονικοποίηση χαρακτήρων (lowercasing), διαγραφή σημείων στίξης. Το κείμενο γίνεται διακριτικό (tokenized) και σχολιάζεται με μέρος των ετικετών ομιλίας (part of

speech tags), ένα βήμα προεπεξεργασίας που απαιτείται για να ενεργοποιηθεί η εφαρμογή συντακτικών φίλτρων. Για να αποφευχθεί η υπερβολική αύξηση του μεγέθους του γραφήματος λόγω της προσθήκης όλων των πιθανών συνδυασμών ακολουθιών πολλών λέξεων (n-grams), το σύστημα επιλέγει μόνο μεμονωμένες λέξεις ως υποψήφιες για το γράφημα. Οι πολλές λέξεις-κλειδιά δημιουργούνται στη φάση της μετά-επεξεργασίας.

Έπειτα, όλες οι λεξικές μονάδες που πληρούν τα συντακτικά κριτήρια προστίθενται στο γράφημα, ενώ προστίθεται μια ακμή μεταξύ εκείνων των λεξικών μονάδων που εμφανίζονται μαζί σε ένα παράθυρο μεγέθους Ν λέξεων. Αφού κατασκευαστεί το γράφημα (μη κατευθυνόμενο και μη σταθμισμένο γράφημα), η βαθμολογία που σχετίζεται με κάθε κορυφή ορίζεται σε μια αρχική τιμή 1 και ο αλγόριθμος κατάταξης εκτελείται στο γράφημα για πολλές επαναλήψεις έως ότου συγκλίνει, συνήθως για 20-30 επαναλήψεις, σε ένα όριο 0,0001. Ο τρόπος λειτουργίας του αλγορίθμου κατάταξης βασίζεται στην ιδέα της “ψηφοφορίας” ή της “σύστασης” μεταξύ των κορυφών ενός γράφου. Όταν μια κορυφή συνδέεται με μια άλλη, ουσιαστικά ψηφίζει υπέρ αυτής της άλλης κορυφής. Η σημασία μιας κορυφής αυξάνεται ανάλογα με τον αριθμό των ψήφων που λαμβάνει. Επιπλέον, το βάρος κάθε ψήφου καθορίζεται από τη σημασία της κορυφής που την παρέχει, λαμβάνοντας υπόψη την αξιολόγηση της κορυφής αυτής. Άρα, η βαθμολογία που σχετίζεται με μια κορυφή καθορίζεται με βάση τις ψήφους που λαμβάνει αυτή η κορυφή και τη βαθμολογία των κορυφών που δίνουν αυτές τις ψήφους. Ως αποτέλεσμα, όλες οι λεξικές μονάδες που πληρούν τα συντακτικά κριτήρια προστίθενται στο γράφημα. Στη συνέχεια, για κάθε ζεύγος λεξικών μονάδων που βρίσκονται μαζί σε ένα παράθυρο Ν λέξεων, προστίθεται μια ακμή. Αρχικά, κάθε κορυφή λαμβάνει μια αρχική βαθμολογία 1, και στη συνέχεια εκτελείται ο αλγόριθμος κατάταξης.

Μετά τον υπολογισμό της τελικής βαθμολογίας για κάθε κορυφή στο γράφημα, οι κορυφές ταξινομούνται κατά φθίνουσα σειρά βαθμολογίας. Στη συνέχεια, διατηρούνται οι κορυφές με τις υψηλότερες βαθμολογίες για περαιτέρω επεξεργασία. Κατά τη μετα-επεξεργασία, όλες οι λεξιλογικές μονάδες που επιλέγονται ως πιθανές λέξεις-κλειδιά, επισημαίνονται στο κείμενο και οι ακολουθίες γειτονικών λέξεων-κλειδιών συμπτύσσονται σε μια λέξη-κλειδί πολλών λέξεων. Οι κορυφές με τις υψηλότερες βαθμολογίες κατατάσσονται ως λέξεις-κλειδιά και επιστρέφονται ως τα σημαντικότερα λεξικά στοιχεία του κειμένου.

3.2.1 Υλοποίηση στην Python

Η υλοποίηση στην python έχει γίνει με την βιβλιοθήκη ‘spaCy’ και ‘PyTextRank’ τα οποία περιέχουν έτοιμη την υλοποίηση του αλγορίθμου TextRank που περιγράφεται

παραπάνω και εμείς απλώς την χρησιμοποιούμε. Ο τρόπος με τον οποίο το χρησιμοποιούμε φαίνεται στην παρακάτω εικόνα.

```
# Load the spaCy model
nlp = spacy.load('/home/kali/Desktop/pythonProject/venv/Lib/site-packages/en_core_web_sm/en_core_web_sm-3.6.0/')

# add PyTextRank to the spaCy pipeline
nlp.add_pipe("textrank")
doc = nlp(extracted_text)

# Examine the top-ranked phrases in the document
temp_textrank = []
for phrase in doc._.phrases:
        # Check if the length of the phrase is 2 or less
        if len(phrase.text.split()) <= max_length:
                temp_textrank.append([phrase.text, phrase.rank])
```

Αυτό το τμήμα κώδικα χρησιμοποιεί το spaCy και το PyTextRank για την εξαγωγή και την αξιολόγηση φράσεων από ένα κείμενο. Αυτός ο κώδικας εκτελεί τα εξής βήματα:

- Φορτώνει το μοντέλο spaCy που έχει εγκατασταθεί στο σύστημά.
- Προσθέτει το PyTextRank στο pipeline του spaCy.
- Αναλύει το κείμενο χρησιμοποιώντας το μοντέλο spaCy.
- Εξετάζει τις φράσεις με την υψηλότερη βαθμολογία στο έγγραφο.
- Αποθηκεύει τις φράσεις που έχουν μήκος όσο θέλουμε να είναι το μέγεθος των λέξεων κλειδιών, μαζί με τη βαθμολογία τους.

3.3 Term Frequency-Inverse Document Frequency

Το Term Frequency-Inverse Document Frequency (TF-IDF) είναι το αρχικό του Όρος Συχνότητας-Αντίστροφης Συχνότητας Εγγράφου [60]. Πρόκειται για ένα αριθμητικό στατιστικό μέτρο που αντικατοπτρίζει τη σημασία ενός όρου (λέξης ή φράσης) σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων (corpus). Είναι πολύ δημοφιλής στην εξαγωγή λέξεων-κλειδιών και τον αυτόματο εντοπισμό θεμάτων σε μεγάλα σύνολα κειμένων λόγω της αποτελεσματικότητάς του στην αντιμετώπιση του προβλήματος της σημασιολογικής συνάφειας.

Αρχικά, υπολογίζει την συχνότητα εμφάνισης κάθε λέξης στο κείμενο, γνωστή και ως Term Frequency (TF). Αυτό απλά αντιπροσωπεύει πόσο συχνά εμφανίζεται μια λέξη σε ένα κείμενο.

$$TF = \frac{number\ of\ the\ term\ appears\ \text{in}\ the\ document}{total\ number\ of\ terms\ \text{in}\ the\ document}$$

Στη συνέχεια, υπολογίζεται η Αντίστροφη Συχνότητα Εγγράφων (Inverse Document Frequency - IDF). Αυτός ο δείκτης μετράει τη σπανιότητα μιας λέξης σε σχέση με το σύνολο των εγγράφων ή το συγκεκριμένο κείμενο. Συνήθως, χρησιμοποιείται ο

λογάριθμος του συνόλου των εγγράφων διαιρούμενος με τον αριθμό των εγγράφων που περιέχουν τη συγκεκριμένη λέξη, για να προσδιορίσει τη σημαντικότητά της.

$$IDF = \log\left(\frac{number\ of\ the\ documents\ \text{in}\ the\ corpus}{number\ of\ documents\ \text{in}\ the\ corpus\ that\ cointain\ the\ term}\right)$$

Τελικά, ο υπολογισμός του TF-IDF γίνεται πολλαπλασιάζοντας τη συχνότητα του όρου (TF) με την αντίστοιχη αντίστροφη συχνότητα εγγράφου (IDF). Αυτός ο υπολογισμός δίνει υψηλό βαθμό σε λέξεις που εμφανίζονται συχνά σε ένα συγκεκριμένο κείμενο αλλά σπάνια σε άλλα κείμενα.

$$TF - IDF = TF * IDF$$

Τέλος, οι λέξεις με τα υψηλότερα σκορ TF-IDF θεωρούνται ως λέξεις-κλειδιά και επιλέγονται για χρήση στην αναπαράσταση του σημασιολογικού περιεχομένου του κειμένου.

Υπάρχουν διάφορες προσεγγίσεις για τον υπολογισμό του IDF σκορ. Συχνά χρησιμοποιείται ο λογάριθμος βάσης 10 για τον υπολογισμό του, ενώ ορισμένες βιβλιοθήκες χρησιμοποιούν φυσικό λογάριθμο. Επιπλέον, προστίθεται ένας μικρός αριθμός στον παρονομαστή για να αποφευχθεί η διαίρεση με το μηδέν.

$$IDF = \log\left(\frac{number\ of\ the\ documents\ \text{in}\ the\ corpus}{number\ of\ documents\ \text{in}\ the\ corpus\ that\ contain\ the\ term + 1}\right)$$

Το TF-IDF είναι ένα εργαλείο που χρησιμοποιείται σε πολλές εφαρμογές επεξεργασίας φυσικής γλώσσας. Μεταξύ άλλων, οι μηχανές αναζήτησης χρησιμοποιούν το TF-IDF για να αξιολογήσουν τη σημασία ενός εγγράφου σε σχέση με ένα ερώτημα. Επιπλέον, το TF-IDF εφαρμόζεται σε διάφορες διαδικασίες όπως η ταξινόμηση κειμένων, η δημιουργία συνόψεων κειμένου και η μοντελοποίηση θεμάτων.

3.3.1 Υλοποίηση στην Python

Η υλοποίηση στην python έχει γίνει με την βιβλιοθήκη ‘TfidfVectorizer’ και ‘pandas’ τα οποία περιέχουν έτοιμη την υλοποίηση του αλγορίθμου TF-IDF που περιγράφεται παραπάνω και εμείς απλώς την χρησιμοποιούμε. Ο τρόπος με τον οποίο το χρησιμοποιούμε φαίνεται στην παρακάτω εικόνα.

```
1  # Create the TF-IDF vectorizer
2 tfidf_vectorizer = TfidfVectorizer(ngram_range=(min_length, max_length), stop_words='english')
3
4 # Fit the vectorizer to the text and transform the text into a TF-IDF vector
5 tfidf_vector = tfidf_vectorizer.fit_transform([extracted_text])
6
7 # Get the feature names (words)
8 feature_names = tfidf_vectorizer.get_feature_names_out()
9 tfidf_scores = tfidf_vector.toarray()[0]
```

Αυτό το τμήμα κώδικα χρησιμοποιεί τον TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer για τον υπολογισμό του TF-IDF πίνακα για ένα κείμενο. Αυτός ο κώδικας εκτελεί τα εξής βήματα:

- Δημιουργεί έναν TF-IDF vectorizer με τις παραμέτρους που έχουν οριστεί.
- Προσαρμόζει τον vectorizer στο κείμενο και μετατρέπει το κείμενο σε ένα TF-IDF vector.
- Λαμβάνει τα ονόματα των χαρακτηριστικών (λέξεων) χρησιμοποιώντας τη μέθοδο get_feature_names_out().
- Λαμβάνει τα TF-IDF σκορ χρησιμοποιώντας τη μέθοδο toarray().

4. Αλγόριθμος RAKE με Σημασιολογία

Σε έναν κόσμο που εξαρτάται όλο και περισσότερο από τις τεχνολογίες επεξεργασίας φυσικής γλώσσας (NLP), η αναζήτηση για ακρίβεια και αποτελεσματικότητα στην ανάλυση κειμένου δεν ήταν ποτέ πιο κρίσιμη. Η ταυτοποίηση συνωνύμων, μια θεμελιώδης εργασία στο NLP, παίζει καθοριστικό ρόλο σε διάφορες εφαρμογές όπως οι μηχανές αναζήτησης, τα συστήματα συστάσεων και η ανάλυση συναισθημάτων. Ως απάντηση σε αυτή την απαίτηση, παρουσιάζουμε μια νέα λύση που αξιοποιεί τη δύναμη των αλγορίθμων Rake και Word2Vec (GloVe).

4.1 Rapid Automatic Keyword Extraction

Ο αλγόριθμος Rapid Automatic Keyword Extraction (RAKE) [56] είναι μια μέθοδος που χρησιμοποιείται για την αυτόματη εξαγωγή λέξεων-κλειδιών ή φράσεων-κλειδιών από ένα σώμα κειμένου. Αναπτύχθηκε ειδικά για εργασίες επεξεργασίας φυσικής γλώσσας, όπως η σύνοψη κειμένου, η ανάκτηση πληροφοριών και η εξαγωγή πληροφοριών από κείμενο. Βασίζεται σε γραφήματα η οποία χρησιμοποιεί συν-εμφανίσεις, που χρησιμοποιεί τόσο τη συχνότητα λέξης όσο και τον βαθμό λέξης για να εκχωρήσει βαθμολογίες σε φράσεις. Λαμβάνει ως παραμέτρους εισαγωγής μια λίστα λέξεων τερματισμού, ένα σύνολο οριοθετημένων φράσεων και ένα σύνολο οριοθετημένων λέξεων για να χωρίσει το κείμενο σε υποψήφιες φράσεις. Στη συνέχεια, δημιουργείται μια γραφική παράσταση με τις συνυπάρχουσες λέξεις και μια βαθμολογία (συχνότητα λέξης ή ο βαθμός λέξης ή η αναλογία βαθμού προς συχνότητα) για κάθε υποψήφια φράση που είναι το άθροισμα των βαθμολογιών των λέξεων που αποτελούν το φράση. Επιπλέον, η RAKE είναι σε θέση να ανιχνεύει φράσεις-κλειδιά που περιλαμβάνουν εσωτερικές λέξεις-κλειδιά. Αυτό γίνεται εντοπίζοντας ζεύγη λέξεων που εμφανίζονται τουλάχιστον δύο φορές με την ίδια σειρά στο ίδιο έγγραφο. Τέλος, οι φράσεις που βρίσκονται στην κορυφή της λίστας T επιλέγονται ως βασικές φράσεις για το έγγραφο.

Ο RAKE ξεκινά την εξαγωγή λέξεων-κλειδιών σε ένα έγγραφο αναλύοντας το κείμενό του σε ένα σύνολο υποψήφιων λέξεων-κλειδιών. Αρχικά, το κείμενο του εγγράφου διαιρείται σε μια σειρά λέξεων βάσει των οριοθετημένων λέξεων. Στη συνέχεια, αυτή η σειρά λέξεων χωρίζεται σε ακολουθίες συνεχόμενων λέξεων, χρησιμοποιώντας οριοθέτες φράσεων και θέσεις λέξεων τερματισμού. Οι λέξεις που ανήκουν σε μια ακολουθία έχουν την ίδια θέση στο κείμενο και θεωρούνται υποψήφιες λέξεις-κλειδιά.

Αφού προσδιοριστεί κάθε υποψήφια λέξη-κλειδί και ολοκληρωθεί το γράφημα των συν-εμφανίσεων λέξεων, υπολογίζεται μια βαθμολογία για κάθε υποψήφια λέξη-κλειδί και ορίζεται ως το άθροισμα των βαθμολογιών των λέξεων-μελών της. Υπολογίζονται διάφορες μετρήσεις για την εκτίμηση των βαθμολογιών λέξεων, με βάση τον βαθμό και τη συχνότητα των κορυφών λέξεων στο γράφημα:

1. Συχνότητα Λέξης (frequency(w)): Αναπαριστά πόσο συχνά εμφανίζεται η λέξη.
2. Βαθμός Λέξης (deg(w)): Αντιπροσωπεύει τον αριθμό των συνδέσεων της λέξης στο γράφημα.
3. Αναλογία Βαθμού προς Συχνότητα (deg(w)/freq(w)): Αυτή η μετρική δίνει έμφαση στις λέξεις που εμφανίζονται συχνά σε μικρό αριθμό από άλλες λέξεις.

Συνολικά, ο βαθμός της κάθε υποψήφιας λέξης-κλειδιού υπολογίζεται ως το άθροισμα των βαθμολογιών των λέξεων που την απαρτίζουν.

Επειδή ο αλγόριθμος RAKE διαχωρίζει τις πιθανές λέξεις-κλειδιά από τις λέξεις-σταματημού, επομένως οι λέξεις-κλειδιά που προκύπτουν δεν περιλαμβάνουν εσωτερικές λέξεις-σταματημού. Παρότι ο αλγόριθμος RAKE έχει προσελκύσει το ενδιαφέρον λόγω της ικανότητάς του να επιλέγει πολύ συγκεκριμένη ορολογία, έχει εκφραστεί ενδιαφέρον για τον εντοπισμό λέξεων-κλειδιών που περιέχουν εσωτερικές λέξεις-σταματημού, όπως ο “άξονας του κακού”. Προκειμένου να εντοπίσει τις λέξεις-κλειδιά, το RAKE αναζητά ζεύγη λέξεων που εμφανίζονται τουλάχιστον δύο φορές στο ίδιο έγγραφο με την ίδια σειρά. Στη συνέχεια, δημιουργείται μια νέα υποψήφια λέξη-κλειδί συνδυάζοντας αυτά τα ζεύγη λέξεων και τις εσωτερικές τους λέξεις-σταματημού. Η βαθμολογία για τη νέα λέξη-κλειδί υπολογίζεται ως το άθροισμα των βαθμολογιών των λέξεων-κλειδιών που την αποτελούν. Επειδή αυτές οι λέξεις-κλειδιά πρέπει να εμφανίζονται δύο φορές με την ίδια σειρά στο έγγραφο, η εξαγωγή τους είναι πιο συνηθισμένη σε μεγάλα κείμενα παρά σε σύντομες περιλήψεις.

Αφού βαθμολογηθούν οι υποψήφιες λέξεις-κλειδιά, οι υποψήφιες λέξεις με την κορυφαία βαθμολογία T συνήθως επιλέγονται ως λέξεις-κλειδιά για το έγγραφο. Το T υπολογίζεται ως το ένα τρίτο του αριθμού των λέξεων στο γράφημα. Διαφορετικά επιστρέφονται όλες οι λέξεις κλειδιά.

Algorithm1: RAKE Algorithm

Input: Research paper in text format
Output: extracted keywords

```
Begin
for every word not in stopwords
        phraselist [] =word
        calculate word scores for phraselist
end for
for every word in phraselist
        compute wordlist_degree=length (word)-1
        then ,  compute word_frequency[word]+=1
endfor
for every word in word_frequency
        computecandidate_word_degree+= word_frequency[word]
endfor
then sort the keywords
return keywords
```

Εικόνα 1. Rake Pseudocode

Πηγή: https://link.springer.com/chapter/10.1007/978-981-10-8636-6_40

4.2 RAKE με Σημασιολογία

Ένας τρόπος να βελτιώσουμε τον αλγόριθμο RAKE είναι να ενσωματώσουμε σημασιολογικές πληροφορίες χρησιμοποιώντας word embeddings όπως το Word2Vec. Με αυτόν τον τρόπο προσφέρεται μια πιο ολοκληρωμένη και σημασιολογικά πλούσια εξαγωγή λέξεων-κλειδιών. Στο πλαίσιο της ανάλυσης σημασιολογίας κειμένου, αποφασίσαμε να χρησιμοποιήσουμε το Word2Vec επειδή υπερέχει στην καταγραφή λεπτομερών σημασιολογικών σχέσεων προβλέποντας λέξεις περιβάλλοντος από μια λέξη-στόχο. Επιλέξαμε αυτόν τον αλγόριθμο έναντι άλλων, όπως το GloVe, επειδή μπορεί να καταγράψει τόσο συντακτικές όσο και σημασιολογικές σχέσεις μεταξύ των λέξεων και είναι ευέλικτο ως προς την προσαρμογή του μεγέθους του παραθύρου περιβάλλοντος, της αρνητικής δειγματοληψίας κ.λπ.

4.2.1 Word2Vec

Το Word2Vec [62] είναι μια δημοφιλής τεχνική επεξεργασίας φυσικής γλώσσας (NLP) που χρησιμοποιείται για την εκμάθηση κατανεμημένων αναπαραστάσεων λέξεων σε έναν συνεχή διανυσματικό χώρο [65]. Εισήχθη από μια ομάδα ερευνητών της Google το

2013. Η θεμελιώδης ιδέα πίσω από το Word2Vec είναι να συλλάβει τη σημασιολογική σημασία των λέξεων αναπαριστάνοντάς τες ως πυκνά διανύσματα, γνωστά και ως ενσωματώσεις λέξεων, με τέτοιο τρόπο ώστε παρόμοιες λέξεις να είναι κοντά σε κάθε μία άλλο στον διανυσματικό χώρο.

Το Word2Vec έχει πολλά πλεονεκτήματα. Είναι ιδιαίτερα αποτελεσματικό στην αναγνώριση των σημασιολογικών σχέσεων μεταξύ των λέξεων, επιτρέποντας τη δημιουργία πλούσιων λεξιλογικών αναπαραστάσεων. Χρησιμοποιείται ευρέως σε εφαρμογές όπως η ανάλυση συναισθήματος, η αναζήτηση πληροφοριών, η μηχανική μετάφραση και πολλά άλλα. Ένα άλλο πλεονέκτημα του Word2Vec είναι η δυνατότητα χρήσης προκαταρκτικών μοντέλων που έχουν εκπαιδευτεί σε τεράστια σύνολα δεδομένων, όπως το Google News. Αυτά τα μοντέλα μπορούν να χρησιμοποιηθούν άμεσα, επιτρέποντας την εξοικονόμηση χρόνου και πόρων.

Στην Python, το Word2Vec μπορεί εύκολα να εφαρμοστεί χρησιμοποιώντας τη βιβλιοθήκη gensim, η οποία παρέχει μια αποτελεσματική υλοποίηση του Word2Vec μαζί με άλλους αλγόριθμους NLP. Επιπλέον, προεκπαιδευμένα μοντέλα Word2Vec που έχουν εκπαιδευτεί σε μεγάλα σώματα, όπως οι Ειδήσεις της Google, η Wikipedia ή το Common Crawl, είναι άμεσα διαθέσιμα για χρήση. Αυτά τα προεκπαιδευμένα μοντέλα μπορούν να φορτωθούν και να χρησιμοποιηθούν απευθείας ή να βελτιστοποιηθούν σε δεδομένα για συγκεκριμένο τομέα, εάν είναι απαραίτητο.

Στην καρδιά του Word2Vec βρίσκεται η ιδέα της μετατροπής λέξεων σε διανύσματα υψηλής διάστασης. Αυτά τα διανύσματα, γνωστά ως ενσωματώσεις λέξεων (word embeddings), αντιπροσωπεύουν τις λέξεις σε έναν συνεχή χώρο όπου οι λέξεις με παρόμοιο νόημα βρίσκονται κοντά μεταξύ τους. Το Word2Vec καταφέρνει να αποτυπώσει τις σημασιολογικές και συντακτικές σχέσεις μεταξύ λέξεων μέσω δύο βασικών μοντέλων. Το Continuous Bag of Words (CBOW) και το Skip-gram [64]. Τα μοντέλα αυτά έχουν την ικανότητα να συλλαμβάνουν τις σημασιολογικές σχέσεις μεταξύ των λέξεων και να τις τοποθετούν σε έναν συνεχή διανυσματικό χώρο.

Το CBOW προσπαθεί να προβλέψει μια λέξη με βάση το περιβάλλον της. Για παράδειγμα, αν έχουμε τη φράση “the quick brown fox jumps over the lazy dog”, το μοντέλο CBOW θα προσπαθήσει να προβλέψει τη λέξη “fox” με βάση τις λέξεις “the quick brown” και “jumps over the lazy dog”. Το CBOW είναι ιδιαίτερα αποτελεσματικό όταν έχουμε ένα μεγάλο σύνολο δεδομένων, καθώς καταφέρνει να μάθει τις σχέσεις μεταξύ των λέξεων με μεγαλύτερη ακρίβεια.

Το Skip-Gram, από την άλλη πλευρά, επιχειρεί να κάνει το αντίθετο. Προβλέπει τις λέξεις του περιβάλλοντος για μια δεδομένη λέξη. Για παράδειγμα, αν έχουμε τη λέξη “fox”, το μοντέλο θα προσπαθήσει να προβλέψει τις λέξεις “the”, “quick”, “brown”, “jumps”, “over”, “the”, “lazy”, και “dog”. Το Skip-Gram είναι πιο αποτελεσματικό όταν έχουμε ένα μικρότερο σύνολο δεδομένων ή όταν οι λέξεις είναι σπάνιες, καθώς

καταφέρνει να μάθει καλύτερα τις συσχετίσεις μεταξύ των λέξεων ακόμα και με λιγότερα δεδομένα.

Η εκπαίδευση του Word2Vec γίνεται με τη χρήση ενός νευρωνικού δικτύου δύο στρωμάτων. Το πρώτο στρώμα είναι το στρώμα εισόδου και το δεύτερο είναι το στρώμα εξόδου. Η εκπαίδευση γίνεται με τη μέθοδο της ολίσθησης παραθύρου (sliding window). Για κάθε λέξη στο κείμενο, ένα παράθυρο προκαθορισμένου μεγέθους (π.χ. 5 λέξεις πριν και 5 λέξεις μετά) κινείται κατά μήκος του κειμένου, και τα δεδομένα εισάγονται στο μοντέλο.

Επίσης το Word2Vec βασίζεται σε δύο βασικές μεθόδους βελτιστοποίησης: την αρνητική δειγματοληψία (Negative Sampling) και τη Softmax ιεραρχία (Hierarchical Softmax). Αυτές οι μέθοδοι βοηθούν στη βελτιστοποίηση του μοντέλου και στη μείωση του υπολογιστικού κόστους.

1. Αρνητική Δειγματοληψία (Negative Sampling): Αντί να ενημερώσει τα βάρη του μοντέλου για κάθε πιθανή λέξη στο λεξικό, το μοντέλο ενημερώνει μόνο τα βάρη για ένα μικρό δείγμα “αρνητικών” λέξεων που δεν είναι στο τρέχον πλαίσιο.

2. Softmax Ιεραρχία (Hierarchical Softmax): Χρησιμοποιεί ένα δυαδικό δέντρο για να αναπαραστήσει την κατανομή πιθανοτήτων όλων των λέξεων στο λεξικό, επιτρέποντας στο μοντέλο να υπολογίσει τις πιθανότητες πολύ πιο γρήγορα.

Μόλις το Word2Vec εκπαιδευτεί, οι λέξεις αναπαρίστανται ως διανύσματα σε έναν πολυδιάστατο χώρο. Τα διανύσματα αυτά έχουν την ιδιότητα να τοποθετούν τις λέξεις με παρόμοιο νόημα κοντά μεταξύ τους. Για παράδειγμα, οι λέξεις “βασιλιάς” και “βασίλισσα” θα έχουν διανύσματα που βρίσκονται κοντά το ένα στο άλλο, ενώ οι λέξεις “βασιλιάς” και “αυτοκίνητο” θα είναι πιο μακριά. Αυτή η χωρική σχέση επιτρέπει στο Word2Vec να συλλάβει τις σημασιολογικές και συντακτικές σχέσεις μεταξύ των λέξεων. Μια γνωστή ιδιότητα αυτών των διανυσμάτων είναι ότι οι σημασιολογικές σχέσεις μπορούν να μοντελοποιηθούν ως αλγεβρικές πράξεις. Για παράδειγμα, το διάνυσμα της λέξης “βασιλιάς” - το διάνυσμα της λέξης “άντρας” + το διάνυσμα της λέξης “γυναίκα” θα είναι πολύ κοντά στο διάνυσμα της λέξης “βασίλισσα”.

4.2.2 GloVe: Μια εναλλακτική προσέγγιση στο Word2Vec

Το GloVe (Global Vectors for Word Representation) [63] είναι μια άλλη δημοφιλής μέθοδος εκμάθησης ενσωματώσεων λέξεων, που εισήχθη από ερευνητές στο Πανεπιστήμιο του Στάνφορντ το 2014. Παρόμοια με το Word2Vec, το GloVe μαθαίνει επίσης κατανεμημένες αναπαραστάσεις λέξεων σε έναν συνεχή διανυσματικό χώρο. Η βασική ιδέα πίσω από το GloVe είναι η αξιοποίηση παγκόσμιων στατιστικών στοιχείων για την ταυτόχρονη εμφάνιση λέξεων μέσα σε ένα σώμα για την εκμάθηση

ενσωματώσεων λέξεων. Σε αντίθεση με το Word2Vec, το οποίο βασίζεται σε παράθυρα τοπικού περιβάλλοντος, το GloVe χρησιμοποιεί παγκόσμιες στατιστικές ταυτόχρονης εμφάνισης λέξεων-λέξεων για να μάθει ενσωματώσεις λέξεων. Στόχος του είναι να συλλάβει την αναλογία πιθανοτήτων συνεμφάνισης μεταξύ των λέξεων. Αυτό το κάνει ελαχιστοποιώντας μια συνάρτηση απώλειας που “τιμωρεί” τη διαφορά μεταξύ του γινόμενου κουκίδων των διανυσμάτων λέξεων και του λογάριθμου των μετρήσεων συνεμφάνισης. Μόλις εκπαιδευτεί το μοντέλο, το GloVe δημιουργεί ενσωματώσεις λέξεων που αποτυπώνουν τις σημασιολογικές σχέσεις μεταξύ των λέξεων με βάση τις πιθανότητες συν-εμφάνισής τους. Όπως και το Word2Vec το GloVe έχει και αυτό πλεονεκτήματα κάποια από αυτά είναι:

- Παγκόσμιο πλαίσιο: Το GloVe αξιοποιεί παγκόσμιες στατιστικές συν-εμφάνισης λέξης-λέξης, καταγράφοντας όχι μόνο το τοπικό πλαίσιο αλλά και τις παγκόσμιες σημασιολογικές σχέσεις μεταξύ των λέξεων.
- Αποδοτικότητα: Το GloVe είναι αποτελεσματικό στην εκπαίδευση, ακόμη και σε μεγάλα σώματα, και συχνά απαιτεί λιγότερους υπολογιστικούς πόρους σε σύγκριση με άλλες μεθόδους.
- Ενσωματώσεις υψηλής ποιότητας: Οι ενσωματώσεις λέξεων που παράγονται από το GloVe είναι υψηλής ποιότητας και έχουν αποδειχθεί ότι αποδίδουν καλά σε διάφορες εργασίες NLP.

Στην Python, μπορεί να υλοποιηθεί με βιβλιοθήκες όπως το gensim ή το spaCy, ή μπορεί να χρησιμοποιηθούν απευθείας προεκπαιδευμένες ενσωματώσεις GloVe. Προεκπαιδευμένες ενσωματώσεις GloVe είναι διαθέσιμες για λήψη, εκπαιδευμένες σε μεγάλα σώματα όπως το Common Crawl, η Wikipedia και το Twitter.

4.2.3 Online Λεξικό

Επίσης εξετάστηκε η χρήση διαδικτυακών λεξικών για την εύρεση συνωνύμων. Αν και τα διαδικτυακά λεξικά, όπως το Thesaurus.com, παρέχουν άμεσα και αξιόπιστα συνώνυμα, η χρήση τους είναι συχνά περιορισμένη από την ανάγκη για συνεχή σύνδεση στο διαδίκτυο και την έλλειψη προσαρμοστικότητας σε εξειδικευμένα ή πολύπλοκα σύνολα δεδομένων. Επιπλέον, δεν μπορούν πάντα να κατανοήσουν το συμφραζόμενο μιας λέξης με τον ίδιο τρόπο που το κάνει το Word2Vec, γεγονός που καθιστά τις ενσωματώσεις λέξεων πιο ακριβείς και ευέλικτες για πιο σύνθετες εργασίες ανάλυσης κειμένου.

Τα διαδικτυακά λεξικά είναι ψηφιακοί πόροι που παρέχουν ορισμούς, μεταφράσεις, προφορές και άλλες πληροφορίες σχετικά με λέξεις και φράσεις. Μερικά βασικά χαρακτηριστικά είναι οι ορισμοί, συνώνυμα και αντώνυμα, παραδείγματα, προφορά

και ετυμολογία. Επίσης χωρίζονται σε δυο τύπους διαδικτυακών λεξικών μονόγλωσσα και δίγλωσσα λεξικά. Τα μονόγλωσσα λεξικά παρέχουν ορισμούς, συνώνυμα, αντώνυμα και άλλες πληροφορίες για λέξεις σε μία μόνο γλώσσα ενώ τα δίγλωσσα λεξικά παρέχουν μεταφράσεις λέξεων και φράσεων μεταξύ δύο γλωσσών.

4.3 Υλοποίηση RAKE με Σημασιολογία

Η υλοποίηση έγινε σε γλώσσα προγραμματισμού Python στην εφαρμογή pycharm η οποία είναι ένα αποκλειστικό Python ολοκληρωμένο περιβάλλον ανάπτυξης (IDE) που παρέχει ένα ευρύ φάσμα βασικών εργαλείων για προγραμματιστές Python, στενά ενσωματωμένα για τη δημιουργία ενός βολικού περιβάλλοντος για παραγωγική ανάπτυξη Python, web και δεδομένων.

Για την χρήση του Word2Vec, χρησιμοποιήθηκε το προ-εκπαιδευμένο μοντέλο της Google, το οποίο έχει εκπαιδευτεί σε ένα τεράστιο σύνολο δεδομένων από το Google News. Αυτό το προ-εκπαιδευμένο μοντέλο περιέχει ενσωματώσεις 3 εκατομμυρίων λέξεων και φράσεων, με διάσταση 300. Χρησιμοποιώντας αυτό το έτοιμο μοντέλο, εξοικονομούμαι χρόνο και πόρους, ενώ ταυτόχρονα διασφαλίζουμε την υψηλή ποιότητα των ενσωματώσεων λέξεων.

Κατά την ανάπτυξη του αλγορίθμου RAKE με σημασιολογία το κίνητρο ήταν να αναπτυχθεί μια μέθοδος εξαγωγής λέξεων κλειδιών που είναι εξαιρετικά αποτελεσματική με την εύρεση συνωνύμων. Ο αλγόριθμος RAKE με σημασιολογία χωρίζεται σε 5 φάσεις από τις οποίες οι 4 είναι παρόμοιες με τον αρχικό RAKE:

1. Προεπεξεργασία
2. Δημιουργία υποψήφιων λέξεων κλειδιών
3. Βαθμολόγηση λέξεων κλειδιών
4. Εξαγωγή λέξεων κλειδιών
5. Εύρεση συνωνύμων στις λέξεις κλειδιά που έχουν εξαχθεί και αφαίρεση αυτών με την μικρότερη βαθμολογία.

Το πρώτο βήμα προεπεξεργασίας ενός εγγράφου είναι η μετατροπή του σε μια μορφή που μπορεί να επεξεργαστεί από μηχανή, προκειμένου να αναγνωριστούν πιθανοί υποψήφιοι όροι. Αυτό είναι ένα κρίσιμο και σημαντικό βήμα για να εντοπιστούν οι καλύτερες υποψήφιες λέξεις κλειδιά και συνεπώς συμβάλλει στην βελτίωση της αποτελεσματικότητας του αλγορίθμου. Ο τρόπος με τον οποίο γίνεται η προεπεξεργασία περιλαμβάνει:

- Την διαχώριση των προτάσεων σε προτάσεις οπού υπάρχουν σημεία στίξης εκτός από την παύλα (-) και την απόστροφο (') επειδή μια λέξη κλειδί μπορεί είναι της

μορφής “K-Means” η “user's” παράδειγμα η πρόταση “Είναι ταλαντούχος, θα δώσει μια συναυλία αύριο” θα διαχωριστεί σε “Είναι ταλαντούχος” και “θα δώσει μια συναυλία αύριο”. Αυτό υλοποιείτε χρησιμοποιώντας την βιβλιοθήκη ‘re’ της python και την συνάρτηση ‘re.findall(r'\b\w+’\w+\b|\b\w+-\w+\b|\b\w+\b')’.

- Αφαίρεση λέξεων στοπ (stopwords) πχ. “and”, “or”.
- Αφαίρεση της παύλας ακολουθημένη από αλλαγή γραμμής ("-\n") επειδή θέλουμε να παίρνει ως μια λέξη τις λέξεις που διακόπτονται πχ. (υπο-\n λογιστής) ως μια. Αυτό το γίνεται διαγράφοντας τους χαρακτήρες ‘-\n’ με την συνάρτηση ‘replace(‘-\n’, ‘’)’.
- Αφαίρεση της αλλαγής γραμμής για να είναι όλες οι λέξεις στην σειρά ώστε να είναι πιο εύκολη η εξαγωγή των λέξεων κλειδιών. Όπως και στο παραπάνω τρόπο προεπεξεργασίας γίνεται με την συνάρτηση ‘replace (‘\n’, ‘’)’
- Μετατροπή όλων των χαρακτήρων σε πεζά χρησιμοποιώντας την συνάρτηση ‘lower()’.

Ο αρχικός RAKE διαχωρίζει τις προτάσεις σε πρόταση με βάση όλα τα σημεία στίξης και δεν αφαιρεί την παύλα (‘-’) ακολουθημένη από αλλαγή γραμμής (’\n’) η σκέτο την αλλαγή γραμμής.

Στην δεύτερη φάση χρησιμοποιούμαι μια κανονική έκφραση (regular expression) για να ανιχνεύσουμε τις λέξεις που περιέχουν απόστροφο (') εντός τους, λέξεις που περιέχουν παύλα (-) εντός τους και τις λέξεις χωρίς ειδικούς χαρακτήρες εντός τους. Επίσης ταυτόχρονα ελέγχουμε άμα οι λέξεις που έχουν ανιχνευθεί είναι έγκυρες λέξεις. Για να είναι οι λέξεις έγκυρες πρέπει να περιέχουν αλφαριθμητικά πχ “human” η “3G”, η να περιέχουν την παύλα ‘-’ και να μην είναι μόνο αριθμοί. Μόλις γίνει αυτό δημιουργούνται δυο λίστες. Η μια περιέχει όλες τις μονολεκτικές λέξεις κλειδιά και η δεύτερη όλες τις φράσεις κλειδιά οπού το μέγεθος της φράσεις είναι ανάλογο με βάση πόσο ορίστηκε ο μέγιστος αριθμός λέξεων. Οι λέξεις/φράσεις κλειδιά βρίσκονται όπως τον κανονικό RAKE με την μονή διαφορά ότι ο δικός μας ανιχνεύει τις λέξεις με παύλα και απόστροφο εντός τους.

Στο τρίτο βήμα, βαθμολογούνται οι λέξεις κλειδιά που δημιουργήθηκαν όπου η υλοποίηση μας στο τρίτο βήμα δεν διαφέρει από τον αρχικό RAKE. Αρχικά υπολογίζεται η συχνότητα κάθε λέξης η οποία είναι ο συνολικός αριθμός που εμφανίζεται κάθε λέξη, χρησιμοποιώντας την συνάρτηση ‘Counter()’ από την βιβλιοθήκη ‘collections’. Έπειτα υπολογίζεται ο βαθμός της κάθε λέξης ο οποίος υπολογίζεται με τον γράφο συν-εμφάνισης λέξεων. Στον γράφο αυτόν, κάθε λέξη αποτελεί έναν κόμβο, ενώ υπάρχει μια ακμή μεταξύ δύο κόμβων αν οι αντίστοιχες λέξεις εμφανίζονται μαζί σε μια πρόταση. Για παράδειγμα, ας υποθέσουμε το ακόλουθο κείμενο: “The weather today is splendid and the atmosphere is invigorating. Individuals wander cheerfully along the

avenues, relishing the bright day. Kids frolic joyfully in the playgrounds, giggling and shouting with glee. Blossoms flourish all around, infusing the air with their delightful scents. Everyone appears content and brimming with vitality, savoring the splendid nature encircling them. In the suburbs, the meadows are verdant and the leaves rustle softly in the wind. Birds sing harmoniously, creating a calming tune that soothes the spirit. The ambiance is serene and tranquil, and people find the chance to unwind and take pleasure in the stillness". Λόγου του μεγέθους του γράφου συν-εμφανίσεων παίρνουμε το 1/3 του το οποίο θα μοιάζει κάπως έτσι:

	weather	today	splendid	atmosphere	invigorating	individuals	wander	cheerfully
weather	1							
today		1						
splendid			2					
atmosphere				1				
invigorating					1			
individuals						1	1	
wander						1	1	1
cheerfully							1	1

Στον παραπάνω γράφο, οι λέξεις "weather", "today", "splendid", "atmosphere", "invigorating", "individuals", "wander" και "cheerfully" αποτελούν κόμβους, ενώ οι ακμές δείχνουν τις σχέσεις μεταξύ των λέξεων. Για παράδειγμα, η λέξη "individuals" συνδέεται με την λέξη "wander", καθώς εμφανίζεται μαζί του σε προτάσεις του κειμένου. Όπου υπάρχει ακμή αναμεσά σε δυο ίδιες λέξεις είναι η συχνότητα της λέξης ενώ όπου υπάρχει ακμή και δεν είναι η ιδιά λέξη σημαίνει ότι η λέξη εμφανίζεται με αυτή την λέξη. Ο βαθμός των λέξεων βγαίνει αθροίζοντας όλες τις ακμές μιας λέξης, πχ. με το παραπάνω παράδειγμα ο βαθμός της "individuals" είναι 2. Οι βαθμολογίες των λέξεων υπολογίζονται διαιρώντας τον βαθμό με την συχνότητα (deg(w)/freq(w)) όπου 'w' είναι η λέξη. Αυτό γίνεται δημιουργώντας ένα λεξικό χρησιμοποιώντας την κλάση 'defaultdict' από την βιβλιοθήκη 'collections' και υπολογίζονται οι βαθμολογίες. Έπειτα προστίθενται οι βαθμολογίες στις φράσεις, πχ. παραπάνω μια φράση είναι η "individuals wander" οπότε η συνολική βαθμολογία της φράσης αυτής είναι η βαθμολογία της λέξης "individuals" συν την βαθμολογία της λέξης "wander" δηλαδή 4 (βγαίνει τόσο επειδή το deg/freq του "individuals" και "wander" είναι 2(2/1)). Έπειτα επιστρέφονται οι λέξεις/φράσεις κλειδιά με τις βαθμολογίες τους χωρίς όμως να υπάρχουν οι λέξεις που δεν υπάρχουν μονές τους. Δηλαδή η λέξη "cheerfully" δεν

υπάρχει μονή τους όπως η λέξη “atmoshpere”, εμφανίζεται μόνο μαζί με άλλες λέξεις κλειδιά.

Στη συνέχεια, στο τέταρτο βήμα τα αποτελέσματα ταξινομούνται με την συνάρτηση ‘sorted()’ με βάση την υψηλότερη βαθμολογία προς την μικρότερη και ανάλογα την επιλογή του χρήστη επιστρέφονται το 1/3 των λέξεων κλειδιών η όλες οι λέξεις κλειδιά και εξάγονται.

Στο πέμπτο βήμα εφαρμόζουμε το Word2Vec για να βρεθούν τα συνώνυμα με τη συνάρτηση ‘most_similar(“η λέξη”, top_n=‘να επιστρέψει τις κορυφαίες N λέξεις’)’. Αυτή η συνάρτηση επιστρέφει μια λίστα με λέξεις που έχουν υψηλή ομοιότητα με τη δεδομένη λέξη κλειδί. Η λίστα με τα συνώνυμα περιλαμβάνει ζευγάρια (λέξη, βαθμολογία), όπου η βαθμολογία αντιπροσωπεύει την ομοιότητα της κάθε λέξης με την αρχική λέξη κλειδί. Αυτή η ομοιότητα είναι ένα μέτρο του κατά πόσο οι λέξεις είναι κοντά στον διανυσματικό χώρο. Έπειτα κρατάμε σε μια λίστα τα συνώνυμα της κάθε λέξης που υπάρχουν στις εξαχθέντες λέξεις κλειδιά από το βήμα 4. Από αυτές τις λέξεις οποία έχει την μεγαλύτερη βαθμολογία την κρατάμε και προσθέτουμε τις βαθμολογίες των άλλων λέξεων σε αυτήν με την υψηλότερη βαθμολογία και αφαιρούνται οι άλλες λέξεις από τις εξαχθέντες λέξεις κλειδιά. Για παράδειγμα άμα έχουμε την λέξη “good” και τα συνώνυμα της είναι τα “exceptional”, “favorable”, “great” και “marvelous” και το συνώνυμο “great” είναι στις εξαγόμενες λέξεις κλειδιά και έχει πιο υψηλή βαθμολογία από την λέξη “good”. Τότε κρατάμε την λέξη “good” και προσθέτουμε την βαθμολογία της λέξης ‘great’ στην λέξη “good”. Αυτό γίνεται μόνο για τις λέξεις κλειδιά και όχι για τις φράσεις κλειδιά. Τέλος επιστρέφονται οι νέες βαθμολογίες.

5. Αξιολόγηση

Για να αξιολογήσουμε την απόδοση του αλγορίθμου RAKE με σημασιολογία το δοκιμάσαμε έναντι μιας συλλογής σε ολόκληρα τα αρχεία pdf και στο abstract αυτών των αρχείων. Η συλλογή αποτελείται από 211 επιστημονικές δημοσιεύσεις. Οι εξαγόμενες λέξεις κλειδιά για κάθε αρχείο συγκρίνονται με τις λέξεις κλειδιά του συγγραφέα. Το μέγεθος των εξαγομένων λέξεων κλειδιών είναι ανάλογο της μεγαλύτερης σε μέγεθος λέξης κλειδί του συγγραφέα, δηλαδή άμα η μεγαλύτερη σε μέγεθος λέξη κλειδί που έχει εκχωρήσει ο συγγραφέας είναι 3 λέξεις τότε εμείς εξάγουμε τις λέξεις κλειδιά μας με μέγιστο αριθμό λέξεων ανά φράση κλειδί σε 3.

Ο τρόπος αξιολόγησης γίνεται χρησιμοποιώντας την βαθμολογία F1. Η βαθμολογία είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης και αποτελεί καλύτερο μέτρο όπως είναι η ακρίβεια. Το F1 score υπολογίζεται:

$$F1\,Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

όπου το precision:

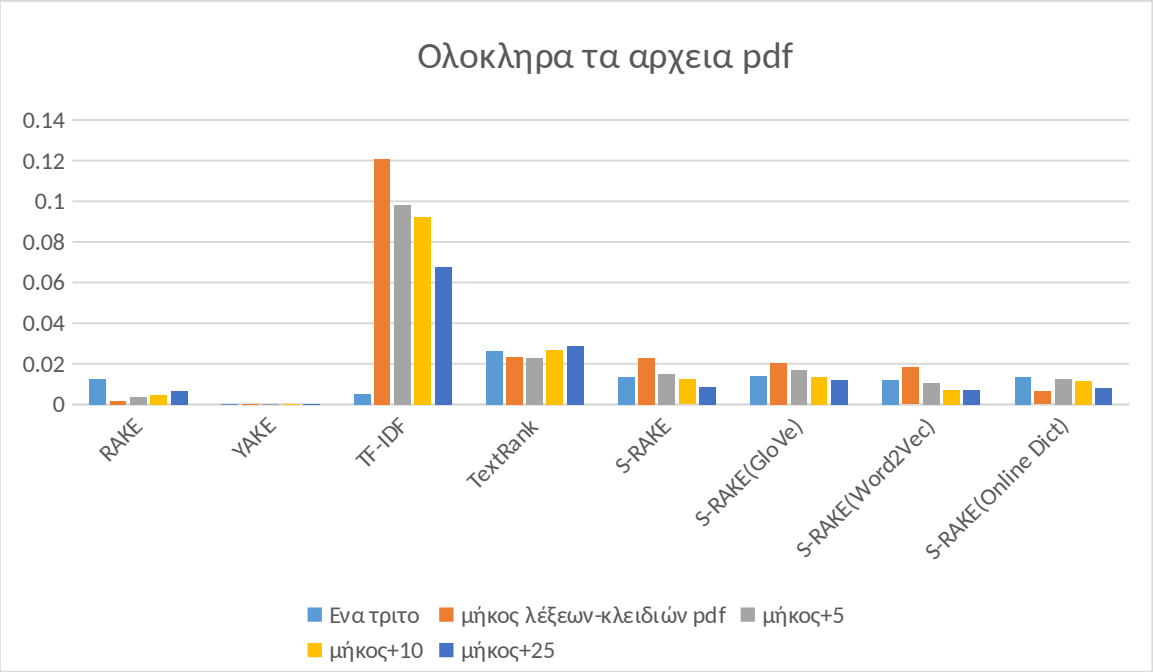
$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

και το Recall:

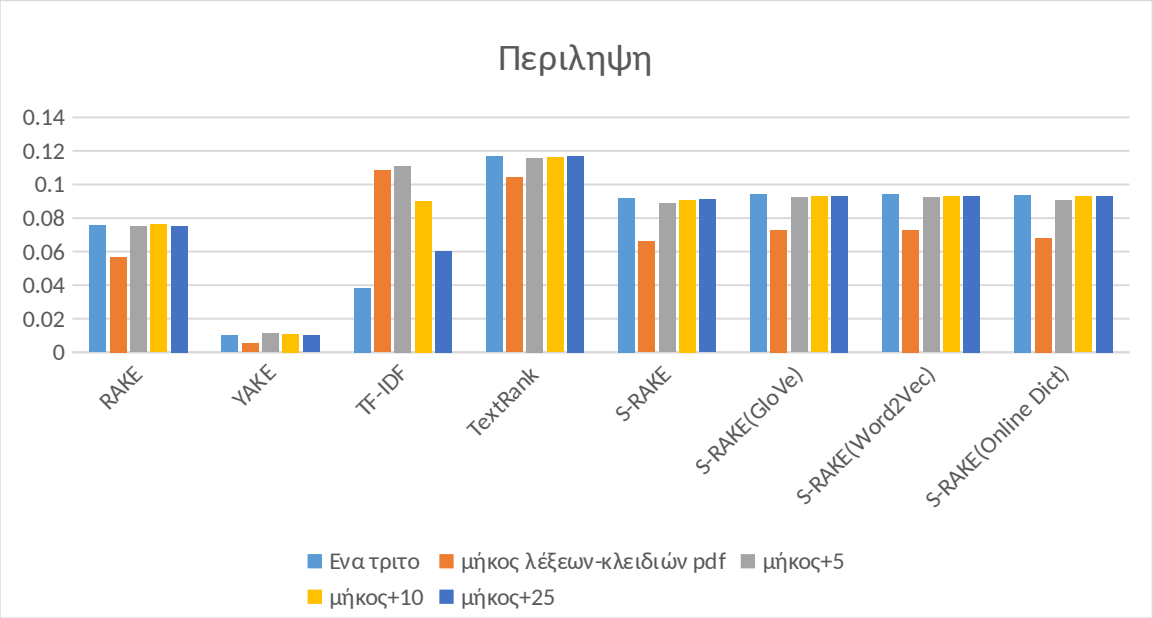
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

Το precision (predictive positive rate) είναι η θετική προγνωστική αξία ενώ το recall (true positive rate) είναι γνωστό ως ευαισθησία (sensitivity) ή αληθώς θετικό ποσοστό. Το True Positive αναφέρεται στην περίπτωση όπου μια λέξη που έχουμε εξάγει αποδεικνύεται πράγματι ως λέξη-κλειδί. Αντίστοιχα, το False Positive αφορά την περίπτωση όπου μια λέξη που έχουμε εξάγει δεν είναι πραγματικά λέξη-κλειδί. Το False Negative είναι όταν δεν έχουμε εξάγει μια λέξη που πράγματι είναι λέξη-κλειδί. Ιδανικά η βαθμολογία F1 θα πρέπει να είναι 1 (υψηλή) για να θεωρείται καλός ταξινομητής.

Στην εφαρμογή μας συγκρίνουμε τον αλγόριθμο RAKE με σημασιολογία (S-RAKE) με τον αλγόριθμο μας χωρίς τα συνώνυμα, με τα συνώνυμα Word2vec, με τα συνώνυμα GloVe, με συνώνυμα από online λεξικά, με το πρωτότυπο RAKE, με το YAKE, με το TF-IDF και με το TextRank.



Εικόνα 2. Αποτελέσματα βαθμολογίας F1 από ολόκληρα τα αρχεία pdf



Εικόνα 3. Αποτελέσματα βαθμολογίας F1 μόνο από την περίληψη

Παραπάνω παρουσιάζονται τα αποτελέσματα της βαθμολογίας F1 από τους αλγορίθμους όπου το “S-RAKE” είναι η υλοποίηση μας του αλγορίθμου RAKE χωρίς την σημασιολογία και τα S-RAKE (GloVe), S-RAKE (Word2Vec) και S-RAKE (Online Dict) είναι η υλοποίηση μας με σημασιολογία. Στην Εικόνα 2. εμφανίζονται τα αποτελέσματα όταν εξετάσαμε ολόκληρα τα αρχεία PDF, ενώ στην Εικόνα 3. εμφανίζονται τα αποτελέσματα όταν χρησιμοποιήσαμε μόνο την περίληψη. Στις εικόνες φαίνονται για κάθε αλγόριθμο οι βαθμολογίες που βρέθηκαν με βάση το ποσοστό των κορυφαίων N λέξεων-κλειδιών που πήραμε. Στο “μήκος λέξεων-κλειδιών pdf” πήραμε τον αριθμό των εξαγόμενων

λέξεων-κλειδιών ίσο με τον αριθμό των λέξεων-κλειδιών στο εξεταζόμενο αρχείο PDF. Δηλαδή, αν οι λέξεις-κλειδιά στο έγγραφο PDF είναι 6, τότε θα πάρουμε τις κορυφαίες 6 λέξεις-κλειδιά. Στα “μήκος + 5”, “μήκος + 10” και “μήκος + 25” προσθέτουμε στα “μήκος λέξεων-κλειδιών pdf” τον αριθμό 5, 10 και 25 αντίστοιχα. Στο “ένα τρίτο” έχουμε πάρει το κορυφαίο ένα τρίτο από όλες τις εξαγόμενες λέξεις-κλειδιά.

Στους παρακάτω πίνακες φαίνονται οι μεσοί οροί από τις βαθμολογίες των αλγορίθμων αναλυτικά με τα precision, recall και F1.

Πίνακας 1. Μέσοι όροι από ολόκληρο αρχείο pdf

	Precision	**Recall**	**F1**
RAKE	0.003423297	0.109928726	0.005761068
YAKE	6.25793E-05	0.0028452	0.000118838
TF-IDF	0.060503528	0.354679843	0.07673722
TextRank	0.016582569	0.225208815	0.025647548
S-RAKE	0.010053606	0.143558715	0.014427282
S-RAKE(GloVe)	0.012518099	0.137994298	0.015345492
S-RAKE(Word2Vec)	0.009985428	0.130271572	0.011012828
S-RAKE(Online Dict)	0.099095487	0.131317235	0.087730358

Πίνακας 2. Μέσοι όροι από τις περιλήψεις

	Precision	**Recall**	**F1**
RAKE	0.092172209	0.098339352	0.071684281
YAKE	0.009126388	0.019200331	0.009472004
TF-IDF	0.088607745	0.173411194	0.081611589
TextRank	0.186171063	0.119180772	0.113917217
S-RAKE	0.096143373	0.130400963	0.085693024
S-RAKE(GloVe)	0.101313139	0.131301437	0.089080536
S-RAKE(Word2Vec)	0.101313139	0.131301437	0.089080536

S-RAKE(Online Dict)	0.006437069	0.119414394	0.010467384

Συγκρίνοντας τους αλγορίθμους, η υλοποίηση μας χωρίς σημασιολογία έχει υψηλότερη βαθμολογία από τον αρχικό RAKE τόσο όταν πρόκειται για ολόκληρο το αρχείο PDF όσο και για τις περιλήψεις.

Όσον αφορά τον αλγόριθμο μας με σημασιολογία, όταν επεξεργάζεται ολόκληρο το αρχείο PDF, η χρήση του GloVe δίνει την καλύτερη βαθμολογία, ακολουθούμενη από το Word2Vec και τα online λεξικά. Ωστόσο, μόνο με τη χρήση του GloVe επιτυγχάνεται καλύτερο αποτέλεσμα από τον αλγόριθμο μας χωρίς σημασιολογία, ενώ με το Word2Vec και τα online λεξικά η βαθμολογία είναι χαμηλότερη από αυτήν του αλγορίθμου χωρίς σημασιολογία.

Στις περιλήψεις, ο αλγόριθμος μας με σημασιολογία χρησιμοποιώντας GloVe και Word2Vec αποδίδει την ίδια βαθμολογία, ενώ ακολουθούν τα online λεξικά. Σε αυτή την περίπτωση, ο αλγόριθμος μας με σημασιολογία υπερβαίνει τον αλγόριθμο χωρίς σημασιολογία.

Όταν συγκρίνουμε τους αλγόριθμους YAKE, TF-IDF και TextRank, παρατηρούμε ότι ο YAKE έχει τη χαμηλότερη βαθμολογία τόσο στις περιλήψεις όσο και στα ολόκληρα αρχεία PDF. Ο TF-IDF επιτυγχάνει την υψηλότερη βαθμολογία σε ολόκληρα αρχεία PDF, με σημαντική διαφορά από τους άλλους αλγόριθμους, αλλά στις περιλήψεις έχει την τρίτη μικρότερη βαθμολογία, ακολουθούμενος από τον κανονικό RAKE και τέλος τον YAKE. Ο TextRank έχει την υψηλότερη βαθμολογία στις περιλήψεις και τη δεύτερη μεγαλύτερη βαθμολογία στα ολόκληρα αρχεία PDF.

Συνολικά, το TextRank και το TF-IDF είχαν τις υψηλότερες βαθμολογίες, ακολουθούμενα από το GloVe τόσο στις περιλήψεις όσο και στα ολόκληρα αρχεία PDF. Μετά από αυτούς τους αλγόριθμους, για ολόκληρα τα αρχεία PDF, η επόμενη καλύτερη βαθμολογία προέρχεται από τον αλγόριθμο μας χωρίς σημασιολογία, ακολουθούμενος από το Word2Vec, τα online λεξικά και τον αρχικο RAKE. Στις περιλήψεις, μετά το GloVe και το Word2Vec, η βαθμολογία μειώνεται για τα online λεξικά, ακολουθούμενα από τον αλγόριθμο μας χωρίς σημασιολογία και τον αρχικο RAKE.

5.1 Αξιολόγηση σε μικρό κείμενο

Εξετάσαμε επίσης τους παραπάνω αλγορίθμους στο ακόλουθο κείμενο για να συγκρίνουμε αν υπάρχει διαφορά όταν το κείμενο που δίνουμε στους αλγορίθμους είναι μεγάλο ή μικρό. Το κείμενο είναι το εξής:

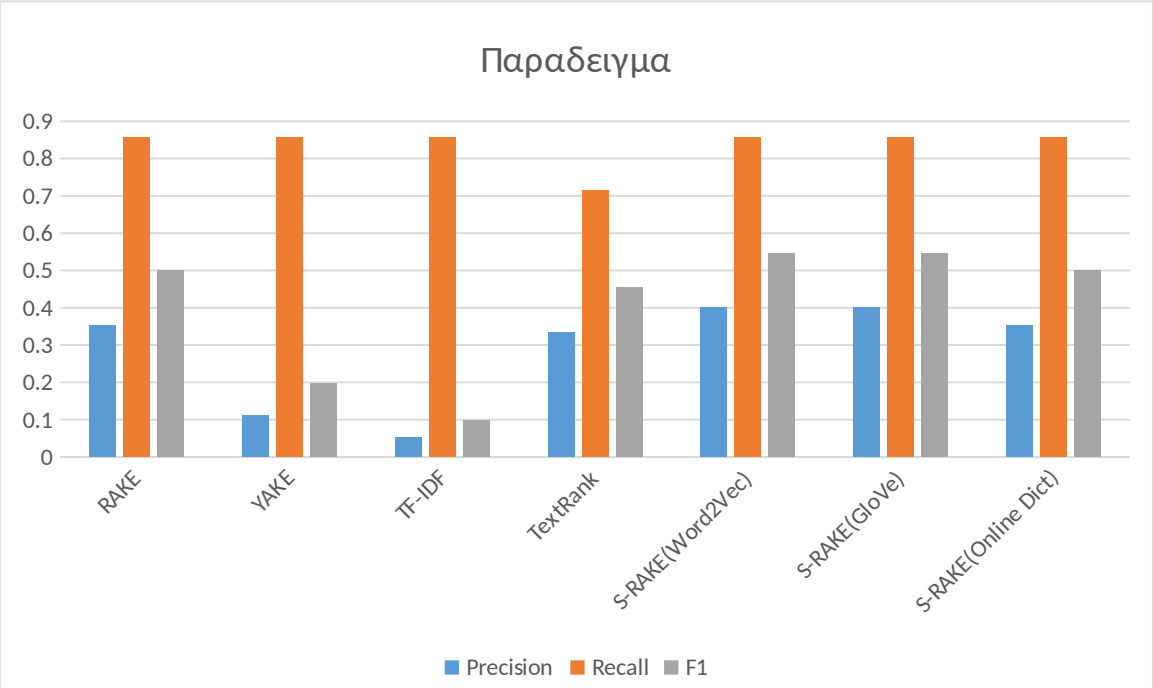
“Compatibility of systems of linear constraints over the set of natural numbers

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal

set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types."

Οι λέξεις-κλειδιά του συγγραφέα είναι: "linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets".

Στην παρακάτω εικόνα παρουσιάζονται τα αποτελέσματα:



Εικόνα 4. Αποτελέσματα του παραδείγματος

Ο αλγόριθμος μας έχει ίδια αποτελέσματα με τον αρχικό RAKE για αυτό και δεν τον αναφέρουμε. Από τα αποτελέσματα της αξιολόγησης προκύπτει ότι οι αλγόριθμοι GloVe και Word2Vec είχαν υψηλότερη βαθμολογία σε σύγκριση με τον αρχικό RAKE, τον RAKE χωρίς σημασιολογία και τα Online λεξικά όπου είχαν την ίδια βαθμολογία. Ο αλγόριθμος TF-IDF κατέγραψε τη χαμηλότερη βαθμολογία, ακολουθούμενος από τον YAKE και τον TextRank. Όσον αφορά το recall, ήταν ίδιο για όλους τους αλγορίθμους εκτός από τον TextRank. Το precision ήταν υψηλότερο όταν χρησιμοποιήθηκαν οι αλγόριθμοι GloVe και Word2Vec, και χαμηλότερο για τους αλγόριθμους YAKE και TF-IDF.

Πίνακας 3. Μέσοι όροι βαθμολογιών από το μικρό κείμενο

	Precision	**Recall**	**F1**

RAKE	0.352941176	0.857142857	0.5
YAKE	0.111111111	0.857142857	0.196721311
TF-IDF	0.052173913	0.857142857	0.098360656
TextRank	0.333333333	0.714285714	0.454545455
S-RAKE(Word2Vec)	0.4	0.857142857	0.545454545
S-RAKE(GloVe)	0.4	0.857142857	0.545454545
S-RAKE(Online Dict)	0.352941176	0.857142857	0.5

Όσο το μοντέλο GloVe όσο και το Word2Vec κατά σύμπτωση δίνουν τα ίδια συνώνυμα για το παραπάνω κείμενο. Βρίσκουν τα συνώνυμα του “systems” και του “constructing” όπου είναι το “system” και το “construction”. Έπειτα διαγράφεται το “constructing” και το “system” και παραμένουν οι άλλες λέξεις/φράσεις κλειδιά. Παρακάτω παρουσιάζεται ο πίνακας με τις λέξεις και τα συνώνυμα τους άλλα μόνο για τις λέξεις όπου υπάρχουν συνώνυμα:

Πίνακας 4. Λέξεις με τα συνώνυμα τους

Λέξη	**Συνώνυμα Word2Vec και GloVe**	**Online λεξικά**
compatibility	compatible , functionality , interoperability , incompatibility , emulation	rapport, unity
systems	system, technologies, devices, software, computer	chip, electronics, wiring
set	setting, sets, up, put, break	
criteria	requirements, criterion, guidelines, standards, maastricht	
system	systems, which, mechanism, control, computerized	arrangement, organization, rule, scheme, structure
considered	regarded, deemed, most, consider, as	studied, treated
components	component, materials, hardware, manufacture, parts	belly, entrails, innards
solutions	solution, solve, strategies, alternatives, innovative	explanation, quick fix, result
algorithms	algorithm, optimization,	

	computation, computational, implementations	
construction	building, projects, project, constructed, built	development, manufacture, plan, planning, structure
types	type, kinds, variety, these, different	brand, breed, category, character, description
constructing	construct, designing, renovating, erecting, construction	
solving	solve, solved, resolving, problem, resolve	corrective, therapeutic

Η σειρά εμφάνισης των λέξεων είναι με βάση την βαθμολογία όπου η πρώτη λέξη έχει την μεγαλύτερη βαθμολογία και εμφανίζονται μόνο οι λέξεις όπου βρέθηκαν συνώνυμα.

Συμπεράσματα

Στην παρούσα εργασία επαυξήσαμε τον αλγόριθμο RAKE προσθέτοντας σημασιολογία μέσω online λεξικών και μοντέλων μηχανικής μάθησης όπως το Word2Vec και το GloVe. Παρόλο που η υλοποίησή μας, τόσο με όσο και χωρίς σημασιολογία, εμφάνισε βελτιωμένα αποτελέσματα σε σύγκριση με τον αρχικό RAKE σε ολόκληρα αρχεία PDF και στις περιλήψεις, οι βαθμολογίες που παρουσιάστηκαν ήταν σχετικά χαμηλές. Ωστόσο, για τη χρήση σε μικρά κείμενα, η υλοποίηση του RAKE με σημασιολογία παρουσιάζει ελαφρώς βελτιωμένα αποτελέσματα από τον αρχικό RAKE, αποδεικνύοντας μια σημαντική βελτίωση στην εξαγωγή λέξεων-κλειδιών. Παρότι αυτή η χρήση σημασιολογίας είχε ως αποτέλεσμα μια μικρή βελτίωση του αλγορίθμου RAKE, παρατηρήσαμε επίσης μείωση της ταχύτητας εκτέλεσης λόγω της πολυπλοκότητας των αλγορίθμων σημασιολογίας.

Η εργασία μελλοντικά μπορεί να επεκταθεί-βελτιωθεί με ποικίλους όπως για παράδειγμα:

- Ανάπτυξη Web Application: Μια από τις προτάσεις είναι η δημιουργία μιας διαδικτυακής εφαρμογής όπου ο χρήστης θα μπορεί να ανεβάζει κείμενα και να λαμβάνει αναλύσεις των σημαντικών λέξεων και φράσεων. Αυτή η εφαρμογή θα μπορούσε να ενσωματώνει τους αλγόριθμους που αναπτύχθηκαν και να προσφέρει μια εύχρηστη διεπαφή για τους χρήστες.
- Χρήση BERT για Εποπτευόμενη Μάθηση: Ένας άλλος τομέας επέκτασης είναι η ενσωμάτωση του αλγορίθμου BERT (Bidirectional Encoder Representations from Transformers) για εποπτευόμενη μάθηση. Ο BERT έχει αποδειχθεί πολύ αποτελεσματικός στην επεξεργασία φυσικής γλώσσας και θα μπορούσε να βελτιώσει την ακρίβεια της εξόρυξης σημαντικών λέξεων και φράσεων.
- Συλλογή Μεγαλύτερου Dataset: Για να πραγματοποιηθεί μια πιο ολοκληρωμένη πειραματική αποτίμηση και αξιολόγηση των αλγορίθμων, είναι απαραίτητο να συγκεντρωθεί ένα μεγαλύτερο και πιο αντιπροσωπευτικό σύνολο δεδομένων. Ένα πλούσιο dataset θα επιτρέψει την καλύτερη κατανόηση των δυνατοτήτων και των περιορισμών των αλγορίθμων, και θα βοηθήσει στη βελτίωση της απόδοσής τους.

Με αυτές τις επεκτάσεις, θα μπορέσουμε να βελτιώσουμε την ακρίβεια και τη χρηστικότητα των εργαλείων που αναπτύξαμε, καθώς και να προσφέρουμε πιο αξιόπιστες και χρήσιμες υπηρεσίες στους τελικούς χρήστες.

Βιβλιογραφία

[1]	D. T. Larose and C. D. Larose, "*Discovering Knowledge in Data: An Introduction to Data Mining*", 2nd ed. Wiley Series on Methods and Applications in Data Mining, D. T. Larose, Ed., Wiley, 2014.
[2]	M. A. North, "*Data Mining for the Masses*", A Global Text Project Book, Creative Commons Attribution 3.0 License, 2012.
[3]	G. Williams, "*Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*", Togaware Pty Ltd, PO Box 655, Jamison Centre, ACT, 2614, Australia, 2011.
[4]	J. Leskovec, A. Rajaraman, and J. D. Ullman, "*Mining of Massive Datasets*", Stanford Univ. and Milliway Labs, 2010-2014.
[5]	M. J. Zaki and W. Meira Jr., "*Data Mining and Analysis: Fundamental Concepts and Algorithms*", 2013.
[6]	D. T. Larose and C. D. Larose, "*Data Mining and Predictive Analysis*", Wiley Series on Methods and Applications in Data Mining, 2015.
[7]	John Wiley & Sons, Inc., "*Data Mining For Dummies*", Hoboken, NJ, 2014.
[8]	D. T. Larose, "*Data Mining: Methods and Models*", John Wiley & Sons, Inc., Hoboken, NJ, 2006.
[9]	M. J. A. Berry and G. S. Linoff, "*Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*", 2nd ed., Wiley Publishing Inc., 2004.
[10]	C. C. Aggarwal, "*Data Mining: The Textbook, Springer*", 2015.
[11]	R. Nisbet, J. Elder, and G. Miner, "*Handbook of Statistical Analysis and Data Mining Applications*", Elsevier, 2009.
[12]	J. Han, M. Kamber, and J. Pei, "*Data Mining*", Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA, 2012.
[13]	I. H. Witten, E. Frank, and M. A. Hall, "*Data Mining: Practical Machine Learning Tools and Techniques*", 3rd ed., Elsevier, 2011.
[14]	S. Stephens-Davidowitz and S. Pinker, "*Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*", Dey St., 2017.
[15]	J. Ledolter, "*Data Mining and Business Analytics with R*", Department of Management Sciences, Tippie College of Business, University of Iowa, Iowa City, Iowa, 2013.
[16]	S. E. Page, "*The Model Thinker*", Basic Books, Hachette Book Group, New York,

	NY, 2018.
[17]	Gaikwad, S.V., Chaugule, A., Patil, P.: "*Text mining methods and techniques*". Int. J. Comput. Appl. 85(17). 2014.
[18]	Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K.: "*A Survey of text mining in social media: facebook and twitter perspectives*". Adv. Sci. Technol. Eng. Syst. J., 2017.
[19]	Navathe, S.B., Ramez, E.: "*Data warehousing and data mining*". Fundam. Database Syst., 841–872,200.
[20]	Gupta, V., Lehal, G.S.: "*A survey of text mining techniques and applications*". J. Emerg. Technol. Web Intell. 1(1), 60–76,2009.
[21]	S. E. Page, "*The Model Thinker*", Basic Books, New York, 2018.
[22]	G. Wiedemann, "*Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*", Springer VS, 2016.
[23]	U. Y. Nahm, "*Text Mining with Information Extraction*", National Science Foundation.
[24]	J. M. G. Hidalgo, "*Text Mining and Internet Content Filtering*", Departamento de Inteligencia Artificial, Universidad Europea CEES.
[25]	D. Pliakuras, "*Text Mining for Drug Discovery*", Manchester, UK, 2014.
[26]	M. Juršič, "*Text Mining for Cross-Domain Knowledge Discovery, Doctoral Dissertation*", Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, May 2013.
[27]	G. Ignatow and R. Mihalcea, "*An Introduction to Text Mining: Research Design, Data Collection, and Analysis*", University of North Texas and University of Michigan, Sage, 2018.
[28]	C. C. Aggarwal and C. Zhai (Eds.), "*Mining Text Data*", Springer, New York Dordrecht Heidelberg London, 2012.
[29]	T. Kwartler, "*Text Mining in Practice with R*", John Wiley & Sons Ltd, 2017.
[30]	J. Hirschberg, E. Hovy, and M. Johnson (Eds.), "*Theory and Applications of Natural Language Processing*", Springer, 2014.
[31]	R. Feldman and J. Sanger, "*The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*", Cambridge University Press, 2007.

[32]	M. W. Berry and M. Castellanos (Eds.), "*Survey of Text Mining II: Clustering, Classification, and Retrieval*", Springer, 2008.
[33]	S. M. Weiss, N. Indurkhya, and T. Zhang, "*Fundamentals of Predictive Text Mining*", Springer, London Heidelberg New York Dordrecht, 2015.
[34]	Pang, Bo and Lee, Lillian. "*Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval*", 2(1-2):1–135, 2008.
[35]	C. C. Aggarwal and C. Zhai, "*Mining Text Data*", Springer, London Heidelberg New York Dordrecht, 2012.
[36]	G.Chakraborty, M. Pagolu, S. Garla, "*Text Mining and Analysis: Practical Methods, Examples and Case Studies Using SAS*", SAS Institute Inc., Cary, NC, USA, 2013.
[37]	J. Silge and D. Robinson, "*Text Mining with R: A Tidy Approach*", O'Reilly, 2017.
[38]	*A. Kongthon, "A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management"*, Georgia Institute of Technology, 2004.
[39]	*A. Kao, S. R. Poteet, "Natural Language Processing and Text Mining", Bellevue, WA98008, USA, 2007.*
[40]	C. P. Chai, *"Statistical Issues in Quantifying Text Mining Performance"*, Department of Statistical Science Duke University, 2017.
[41]	E. Gutierrez, W. Karwowski, K. Fiok, M. R. Davahli, T. Liciaga, and T. Ahram, "*Analysis of Human Behavior by Mining Textual Data: Current Research Topics and Analytical Techniques"* presented at the Symmetry 2021 Conference, Online Conference, July 16, 2021, Paper number 1276.
[42]	J. T. Yun, B. R. L. Duff, P. T. Vargas, H. Sundaram, and I. Himelboim, "*Computationally Analyzing Social Media Text for Topics: A Primer for Advertising Researchers*", December 31, 2019, pp. 47-59.
[43]	J. Goldenstein and P. Poschmann, "*Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling"*, presented at the Sociological Methodology Conference, Germany, 2019, vol. 49, no. 1, pp. 83-131.
[44]	SH. Wang, Y. Ding, W. Zhao, "*Text mining for identifying topics in the literatures about adolescent substance use and depression*", presented at the BMC Public Health Conference, 2016, vol. 16, p. 279.
[45]	D. Korenčić, S. Ristov, J. Repar, and J. Šnajder, "*A Topic Coverage Approach to*

	Evaluation of Topic Models",*q* in IEEE Access, vol. 9, pp. 123280-123312, 2021.
[46]	G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "*Research on Topic Detection and Tracking for Online News Texts*", in IEEE Access, vol. 7, pp. 58407-58418, 2019.
[47]	F. Gurcan and N. E. Cagiltay, "*Exploratory Analysis of Topic Interests and Their Evolution in Bioinformatics Research Using Semantic Text Mining and Probabilistic Topic Modeling*", in IEEE Access, vol. 10, pp. 31480-31493, 2022.
[48]	A. Onan, "*Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering*", in IEEE Access, vol. 7, pp. 145614-145633, 2019.
[49]	R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "*A Survey of Text Representation and Embedding Techniques in NLP*", in IEEE Access, vol. 11, pp. 36120-36146, 2023.
[50]	L. Akhtyamova, P. Martínez, K. Verspoor, and J. Cardiff, "*Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives*", in IEEE Access, vol. 8, pp. 164717-164726, 2020.
[51]	K. Fundel, "*Text Mining and Gene Expression Analysis: Towards Combined Interpretation of High Throughput Data*", Dissertation, Fakultät für Mathematik, Informatik und Statistik, Ludwig–Maximilians–Universität München, Friedrichshafen, April 18, 2007.
[52]	A. Puurula, "*Scalable Text Mining with Sparse Generative Models*", University of Waikato, Feb 8, 2016.
[53]	C. T. Martins, "*A Tool for Text Mining in Molecular Biology Domains*", Departamento de Engenharia Informática, Faculdade de Engenharia da Universidade do Porto, April 2013.
[54]	S. Ananiadou and J. M. Naughton, "*Text Mining for Biology and Biomedicine*", ARTECH HOUSE, INC., Boston, London, 2006.
[55]	S. Siddiqi and A. Sharan, "*Keyword and Keyphrase Extraction Techniques: A Literature Review*", International Journal of Computer Applications, vol. 109, pp. 18-23, Jan. 16, 2015. doi: 10.5120/19161-0607.
[56]	M. W. Berry and J. Kogan, "*Text Mining: Applications and Theory*", John Wiley & Sons, Ltd., 2010.
[57]	S. Beliga, A. Meštrović, and S. Martincic-Ipsic, "*An Overview of Graph-Based*

	Keyword Extraction Methods and Approaches", Journal of Information and Organizational Sciences, vol. 39, pp. 1-20, 2015.
[58]	R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "*YAKE! Keyword Extraction from Single Documents using Multiple Local Features*", Information Sciences, vol. 509, pp. 257-289, 2020. doi: 10.1016/j.ins.2019.09.013.
[59]	R. Mihalcea and P. Tarau, "*TextRank: Bringing Order into Texts*", Department of Computer Science, University of North Texas, 2004.
[60]	S. Qaiser and R. Ali, "*Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents*", Journal Name, vol. 181, no. 1, pp. (pages), July 2018.
[62]	T. Mikolov, K. Chen, G. Corrado, and J. Dean, "*Efficient Estimation of Word Representations in Vector Space*", ArXiv preprint arXiv:1301.3781, 2013.
[63]	J. Pennington, R. Socher, and C. D. Manning, "*Glove: Global Vectors for Word Representation*", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
[64]	T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "*Distributed Representations of Words and Phrases and their Compositionality*", in Advances in Neural Information Processing Systems, 2013.
[65]	Goodfellow, Y. Bengio, and A. Courville, "*Deep Learning*", Cambridge, MA, USA: MIT Press, 2016.

Παράρτημα Κώδικα

Ο πηγαίος κώδικας παρακάτω παρουσιάζει συνοπτικά την δομή του αλγορίθμου S-RAKE και των κυριότερων μεθόδων που έχουν οριστεί σε αυτό.

```
class RakeKeywordExtractor:
    def __init__(self, stop_words_path=None, min_word_length=1, max_words_in_phrase=3,
top_n = False):
        self.stop_words = self.load_stop_words(stop_words_path) if stop_words_path else set()
        self.min_word_length = min_word_length
        self.max_words_in_phrase = max_words_in_phrase
        self.top_n = top_n

def load_stop_words(self, stop_words_path)
def is_valid_word(self, word)
def generate_candidate_keywords(self, sentences)
def find_unique_phrases(self, candidate_phrases)
def find_whole_word(self, word, text)
def find_word(self, word, text)
def keywords_with_more_than_two_words(self, candidate_phrases)
def keywords_with_one_word(self, candidate_phrases)
def calculate_word_scores(self, candidate_phrases, candidate_keywords)
def calc_word_freq(self, candidate_keywords)
def calc_word_degree(self, non_repeated_keyword_list, candidate_phrases)
def rake_extract_keywords(self, text)
def synonyms(self, _keywords_with_scores)
def extract_text_from_pdf(self, pdf_path)

rake = RakeKeywordExtractor(stop_words_path=stop_words, min_word_length=min_length,
max_words_in_phrase=max_length, top_n=False)
extracted_text = rake.extract_text_from_pdf(pdf_path)
keywords_with_scores = rake.rake_extract_keywords(extracted_text)  synonyms =
rake.synonyms(keywords_with_scores)
```