ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ & ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Δημιουργία Συστήματος Αυτόματης Αναγνώρισης Όρων σε Διεθνείς Δημοσιεύσεις και Κλινικές Μελέτες που αφορούν το Σύνδρομο Σιόγκρεν και αξιοποίηση του για ιατρικούς σκοπούς

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρήστου Θ. Βουτσά

Επιβλέπουσα : Θεοδώρα Α. Βαρβαρίγου
Καθηγήτρια ΕΜΠ

Αθήνα, Σεπτέμβριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ & ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Δημιουργία Συστήματος Αυτόματης Αναγνώρισης Όρων σε Διεθνείς Δημοσιεύσεις και Κλινικές Μελέτες που αφορούν το Σύνδρομο Σιόγκρεν και αξιοποίηση του για ιατρικούς σκοπούς

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρήστου Θ. Βουτσά

Επιβλέπουσα : Θεοδώρα Α. Βαρβαρίγου
Καθηγήτρια ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 30 Σεπτεμβρίου 2021.

.....	.....	.....
Θεοδώρα Βαρβαρίγου Καθηγήτρια ΕΜΠ	Εμμανουήλ Βαρβαρίγος Καθηγητής ΕΜΠ	Συμεών Παπαβασιλείου Καθηγητής ΕΜΠ

Αθήνα, Σεπτέμβριος 2021

.....

Χρήστος Θ. Βουτσάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Οι σύγχρονες επιστήμες έχουν μία πληθώρα τεχνικών όρων, με νέους να προστίθενται διαρκώς στους καταλόγους τους. Η χειρωνακτική ενημέρωση των καταλόγων από τους ειδικούς επιστήμονες είναι μια χρονοβόρα διαδικασία που τείνει να γίνει αδύνατη σε λογικά χρονικά πλαίσια. Υπό αυτές τις συνθήκες δημιουργήθηκε η ανάγκη αυτοματοποίησης της διαδικασίας με χρήση υπολογιστικών μεθόδων. Όμοια με το πρόβλημα της Μηχανικής Μετάφρασης τη δεκαετία του ’50, η απόλυτη και εξαντλητική τυποποίηση των κανόνων και των επίσημων γραμματικών στα κείμενα είναι ανεπαρκής για να εντοπιστεί το νόημα στα κείμενα, είτε πρόκειται για μετάφραση, είτε για εξαγωγή όρων. Από την άλλη, η απόλυτη στήριξη στο λεξιλόγιο χωρίς επεξεργασία του κειμένου, δεν αφήνει περιθώρια στην πρόβλεψη των νέων όρων που εισάγονται στην επιστήμη. Για αυτόν τον λόγο, τη δεκαετία του ’90 αναπτύχθηκαν στατιστικές μέθοδοι αναγνώρισης και εξαγωγής των όρων από κείμενα, με στόχο να τυποποιήσουν τη διαδικασία με την οποία μεταβαίνουμε από τη συντακτική και γραμματική ανάλυση του κειμένου στην στατιστική ερμηνεία των λέξεων του. Μία από αυτές ήταν η C-Value, η οποία, για πολυλεκτικές φράσεις, υπολογίζει τον βαθμό ύπαρξης όρου (termhood).

Στην εργασία αυτή παρουσιάζουμε μία εφαρμογή Αυτόματης Αναγνώρισης Όρων που αναπτύχθηκε βασισμένη στις αρχές της C-Value. Επεκτείνοντας τον αλγόριθμο ώστε να συμπεριλαμβάνει και μονολεκτικούς όρους και με την αξιοποίηση βιβλιοθηκών της java για διαδικασίες τοκενοποίησης (tokenization), λεκτικής ανάλυσης, Part-of-Speech Tagging, stemming και κανονικοποίησης των όρων, αντιμετωπίσαμε το συντακτικό σκέλος της ανάλυσης του κειμένου. Στη συνέχεια, όσον αφορά το σημασιολογικό σκέλος, υλοποιήσαμε μία διεπαφή αντιστοίχισης συνωνύμων των ιατρικών όρων με το WordNet, η οποία στα αρχικά πειράματα κρίθηκε ανεπαρκής. Μετά από τις συντακτικές και σημασιολογικές προσεγγίσεις, προβήκαμε σε υπολογισμό του C-Value των ακολουθιών λέξεων που αντιστοιχίζονταν σε μοτίβα Part Of Speech που εισάγαμε. Αυτές οι ακολουθίες θεωρήθηκαν υποψήφιοι όροι και απέκτησαν την τιμή C-Value με βάση έναν μαθηματικό υπολογισμό που συμπεριλάμβανε τις απόλυτες και σχετικές συχνότητες των ιδίων και των υπακολουθιών τους.

Τα αποτελέσματα έδειξαν ότι μία τέτοια εφαρμογή μπορεί να αποτελέσει χρήσιμο εργαλείο υποβοήθησης ενός αναλυτή, καθώς μπορεί να προτείνει με μεγάλη ακρίβεια πραγματικούς ιατρικούς όρους. Όμως, η παρακολούθηση των αποτελεσμάτων από έναν ιατρικό επιβλέποντα κρίνεται απαραίτητη, καθώς κρύβονται και ορισμένα false positives εντός των αποτελεσμάτων. Αυτά εντοπίστηκαν μέσα από την αξιολόγηση που εκτελέσαμε με μικρά datasets αλλά και με πειράματα που έγιναν με χρήση της εφαρμογής για εύρεση όρων σχετικών με το ιατρικό σύνδρομο Sjogren.

Λέξεις Κλειδιά

Αυτόματη Αναγνώριση Όρων, Επεξεργασία Κειμένου, Ιατρικά Άρθρα, Οντολογίες, Σύνδρομο Sjogren

Abstract

Modern science of every branch includes a vast number of technical terms, with new ones continuously enriching their corpora. The manual update of those corpora by specialist scientists is a time-consuming procedure which tends to become impossible to execute in sensible time frames. Under this assumption, a need for automating this procedure with computational methods begun to appear in the middle of the 20th century. Like the Machine Translation problem of the 50s decade, the absolute and exhaustive definition of the formal grammars is inadequate when it comes to detecting the semantics of a text, whether the goal is to translate or to extract terms. On the other hand, sole focus on vocabulary without accounting for syntax, limits our predictions on new terms that will appear in any scientific text. Thus, statistic methods for term recognition and extraction were developed during the 90s, to formalize the procedure of connecting syntax and grammar with statistic measures of the text's contents. C-Value was one of these methods, and it is used to measure the termhood of multi-word phrases.

In this thesis, we present an Automatic Term Extraction application developed under the fundamentals of the C-Value method. By extending the algorithm so that it also considers single-word phrases, and by utilizing java libraries for tokenization, POS-tagging, document preprocessing and term normalization, we achieved the syntactic analysis of texts. As for the semantics, we implemented an interface that detects synonyms of medical terms with the help of WordNet, but our preliminary experiments showed inadequate results. Through our semantics and syntax approaches, we proceeded with the computation of the C-Value of word sequences that corresponded to matching Part Of Speech patterns we provided as input. Those candidate terms' C-Values were calculated based on the absolute and relative frequency not only of their own appearances in the text but also of the appearances of their sub-terms.

Our results show that such an application can be a useful tool of assistance for an analyst, as it can recommend real medical terms with high precision. However, manual oversight of the results by a medical expert is considered necessary, as some false positives exist within the results. Those were tracked by the execution of an evaluation algorithm, using controlled datasets but this observation also happened during our experiments, when we used the application to search for terms relevant to the Sjogren's Syndrome disease.

Keywords

Automatic Term Recognition, Text Processing, Medical Articles, Ontologies, Sjogren Syndrome

Περιεχόμενα

Σχήματα

Πίνακες

Πρόλογος

Στα πλαίσια της Διπλωματικής μου εργασίας για την περαίωση των σπουδών μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (ΗΜΜΥ) ασχολήθηκα με τη μελέτη του κλάδου της Υπολογιστικής Γλωσσολογίας και ιδιαίτερα με τις στατιστικές μεθόδους Ανίχνευσης Όρων σε κείμενα. Η εργασία αυτή μου έδωσε τη δυνατότητα να κατανοήσω τις επιμέρους διαδικασίες επεξεργασίας ενός κειμένου, πώς από ανεπεξέργαστες συμβολοακολουθίες διακρίνουμε λέξεις, μέρη του λόγου και τί είδους μετατροπές πρέπει να συντελεστούν επάνω τους ώστε να είναι δυνατή η στατιστική ανάλυση. Μπόρεσα επιπλέον να καταλάβω πώς με τη σωστή χρήση στατιστικών μεθόδων, μπορούμε να εξάγουμε χρήσιμη πληροφορία από μεγάλα μεγέθη κειμένων.

Μου έδωσε επίσης την ευκαιρία να εξασκήσω στην πράξη τις γνώσεις που αποκόμισα στη σχολή σχετικά με τη δημιουργία εφαρμογών σε Java, με παράλληλη διαχείριση σχεσιακής βάσης και ενσωμάτωση ποικιλίας βιβλιοθηκών, προγραμματίζοντας με βάση διαδεδομένα Μοτίβα Σχεδιασμού (Design Patterns) της γλώσσας Java. Επιπλέον, δόθηκε η αφορμή για να μελετήσω με τη σκοπιά ενός μηχανικού έναν εντελώς διαφορετικό κλάδο της επιστήμης, αυτόν της ιατρικής και να αντιληφθώ στην πράξη πώς η εργασία ενός Μηχανικού Υπολογιστών μπορεί να βοηθήσει σε ποικίλα τμήματα της επιστήμης και της ανθρώπινης ζωής.

Η εργασία που παρουσιάζεται στις παρακάτω σελίδες εκπονήθηκε υπό την επίβλεψη της κ. Θεοδώρας Βαρβαρίγου με τη βοήθεια της ομάδας του εργαστηρίου της, την οποία ευχαριστώ πολύ. Ιδιαίτερες ευχαριστίες θα ήθελα να δώσω στον Διδάκτορα Ευθύμιο Κ. Χονδρογιάννη, ο οποίος ήταν υπεύθυνος για την καθοδήγηση μου στην προετοιμασία και τη συγγραφή αυτή της εργασίας, και με βοήθησε με κατανόηση, καίριες παρατηρήσεις και διαρκή παρουσία καθ' όλη τη διάρκεια της διπλωματικής μου.

Θα ήθελα να ευχαριστήσω τις παρέες και τους φίλους μου, πολλοί εκ των οποίων συμφοιτητές, που έκαναν όλο το ταξίδι εντός της σχολής πιο όμορφό.

Επίσης, η φοίτηση στη σχολή των Ηλεκτρολόγων Μηχανικών και η περαίωση της ήταν εφικτή χάρη στη στήριξη από την οικογένεια μου, την οποία θα ήθελα να ευχαριστήσω και στην οποία αφιερώνω την εργασία αυτή.

Χρήστος Θ. Βουτσάς

Σεπτέμβριος 2021

1 Εισαγωγή

Η εκτεταμένη παρουσία του διαδικτύου στη καθημερινότητα και στην επαγγελματική ζωής πολιτών στις περισσότερες ανεπτυγμένες χώρες έχει ωθήσει σε μια ανεπανάληπτη και εκθετική συσσώρευση δεδομένων, πλήθος εκ των οποίων είναι απαραίτητα για κρίσιμες υποδομές της κοινωνίας, όπως για παράδειγμα το εμπόριο. Στο διαδίκτυο στηρίζονται ολοένα και μεγαλύτερα μέρη από το δίκτυο πρωτογενούς παραγωγής (συσκευές παρακολούθησης γεωργικών προϊόντων) ή δευτερογενούς επεξεργασίας και μεταφοράς προϊόντων. (Ενδεικτικά, κολοσσοί διανομής προϊόντων, όπως η Amazon, βασίζονται πρωτίστως στη διαδικτυακή τους πλατφόρμα για εξυπηρέτηση πελατών). Το 2011 υπολογίστηκε ότι συνολικά είχαν παραχθεί, αναμεταδοθεί ή καταναλωθεί συνολικά περί τα 1.8 zettabytes (=10^{21} bytes) [1], ενώ έως το 2018 είχαμε φτάσει στα 33 zettabytes, με βάση τα δεδομένα της International Data Corporation. [2].

Μέσα σε αυτό το πλήθος δεδομένων, αν εξαιρέσουμε το μεγάλο ποσοστό οπτικοακουστικών δεδομένων, κρύβεται ένας θησαυρός από πραγματικά άρθρα και κείμενα, η πλήρης ανάγνωση των οποίων από κάποιο ανθρώπινο μάτι καθίσταται αδύνατη. Ένας άλλος ζωτικός για τις ανθρώπινες κοινωνίες τομέας εκτός από τον εμπορικό, είναι ο ιατρικός, ο οποίος υπολογίζεται ότι σχετίζεται με κοντά στα 1.2 από αυτά τα 33 zettabytes [2]. Σε αυτήν την τεράστια ποσότητα δεδομένων υπάρχει κρυμμένη ιατρική πληροφορία, ορολογία και χρήσιμες παρατηρήσεις, την οποία δύσκολα εντοπίζει οποιοσδήποτε ιατρικός ερευνητής δεν χρησιμοποιήσει κάποιο πρόγραμμα υποβοηθούμενης ανάγνωσης. Παρόλα αυτά, η ιατρική επιστήμη και έρευνα κατορθώνει να διαχειριστεί όλη αυτή την συσσώρευση δεδομένων και συνεχίζει να εξελίσσεται με αλματώδεις ρυθμούς και με πρωτοφανείς δυνατότητες παγκόσμιας συνεργασίας που δεν ήταν ποτέ άλλοτε εφικτές, όπως αποδεικνύεται με τη διεθνή ανταλλαγή γνώσης ιατρικών ερευνητών κατά την πανδημία του κορονοϊού το 2020, με τεράστιες πηγές δεδομένων ελεύθερα διαθέσιμες για ερευνητική και ακαδημαϊκή ανάλυση (για παράδειγμα το Cord-19 Dataset [3]).

Όλη αυτή η δυνατότητα συνεργασίας δεν μπορεί να πιστωθεί αποκλειστικά στην ανάπτυξη των τηλεπικοινωνιών και μόνο. Αν όλος ο όγκος δεδομένων που εισαγόταν καθημερινά δεν μπορούσε να φιλτραριστεί και να εξαχθεί η ουσία από μέσα του, θα είχαμε να κάνουμε με έναν νέο «πύργο της Βαβέλ». Η δυνατότητα φιλτραρίσματος της πληροφορίας και αντιστοίχισης της σε οντολογίες είναι μια κατάκτηση της Επιστήμης Ανάκτησης Πληροφορίας και της Υπολογιστικής Γλωσσολογίας που

εξελίχθηκε περνώντας από αρκετά στάδια με διαφορετικές προσεγγίσεις στις μεθόδους που θα επιτευχθεί αυτή η μοντελοποίηση της χρήσιμης και ουσιαστικής πληροφορίας. Γραμματικές, στατιστικές, λεξιλογικές μέθοδοι και εργαλεία, τεχνικές αναγνώρισης Μέρους του Λόγου (POS-Tagger) είναι λίγα από τα εργαλεία που έχουν αναπτυχθεί προκειμένου να οργανωθεί ο γραπτός λόγος σε μορφή οργανωμένη για τον υπολογιστή, ώστε να εξάγεται ως πληροφορία «εύπεπτη» για τον άνθρωπο.

Σημαντικό μέρος της διαδικασίας αυτής αποτελεί και η Αυτόματη Αναγνώριση Όρων, με ορισμένη μεθοδολογία και αλγορίθμους η οποία γνώρισε μεγάλη ανάπτυξη στα τέλη του 20ού αιώνα μέσα από τις στατιστικές μεθόδους και κατά κάποιον τρόπο αποτελεί και έναν πρόγονο (ή έστω μακρινό συγγενή) των σημερινών εξαιρετικά διαδεδομένων τεχνικών μηχανικής μάθησης. Έναν από αυτούς τους αλγορίθμους, τον επονομαζόμενο C-Value, υλοποιήσαμε σε αυτήν την εργασία για να μελετήσουμε κι εμείς με τη σειρά μας πώς μπορεί να βοηθήσει στην εξαγωγή όρων που σχετίζονται με το ιατρικό τομέα και δη, για το σύνδρομο Σιόγκρεν. Με τον αλγόριθμο C-Value, εντοπίζονται, αξιολογούνται και κατατάσσονται υποψήφιοι πολυλεκτικοί τεχνικοί όροι που υπάρχουν σε ένα κείμενο, με χρήση ενός συνδυασμού προεπεξεργασίας του κειμένου, POS-Tagging, stemming και στατιστικής ανάλυσης. Αποτελεί τη βάση και μέτρο σύγκρισης για πολλούς αντίστοιχους αλγόριθμους με ελαφρώς διαφορετική αντιμετώπιση των μετρικών που λαμβάνουν υπόψιν.

1.1 Στόχος Εργασίας

Στα πλαίσια της εργασίας αυτής, δημιουργήσαμε έναν αλγόριθμο αναγνώρισης όρων (*TermFinder*) βασισμένο στις αρχές του αλγορίθμου C-Value, τον οποίο ακολούθως χρησιμοποιήσαμε για να εντοπίσουμε τους όρους που χρησιμοποιούνται συχνά σε τρείς διαφορετικές πηγές δεδομένων που σχετίζονται άμεσα ή έμμεσα με το Σύνδρομο Σιόγκρεν. Ο αλγόριθμος υλοποιήθηκε σε java και περιλαμβάνει κατάλληλους preprocessors και parsers για τα δεδομένα εισόδου. (όπως tokenizers και POS Taggers από την Stanford CoreNLP βιβλιοθήκη). Ανάλογα με το configuration, χρησιμοποιεί stemmers (PorterStemmer), ελεγκτές ορθογραφίας (Jazzy) και σημασιολογικούς ελεγκτές, οι οποίοι ενισχύουν το C-Value των πραγματικών όρων με αυτό των παραλλαγών τους. Για την καλύτερη διαχείριση των ενδιάμεσων αποτελεσμάτων της επεξεργασίας των δεδομένων χρησιμοποιήσαμε μία σχεσιακή βάση SQLite.

Προκειμένου να αξιολογήσουμε την εφαρμογή μας, δώσαμε για είσοδο ένα μικρό κείμενο στο οποίο είχαμε ήδη εντοπίσει οπτικά «γνωστούς όρους» και εκτελέσαμε τον αλγόριθμο. Αυτός τροφοδότησε τα αποτελέσματα του σε έναν αλγόριθμο αξιολόγησης αποτελεσμάτων, ο οποίος βγάζει συνολικές

στατιστικές πληροφορίες με βάση αυτά. Η γνώση των όρων που θα αναμέναμε να βγουν και το μικρό κείμενο μας έδωσε ευκολία στο debugging της εφαρμογής.

Στη συνέχεια, στα πλαίσια της εργασίας, χρησιμοποιήσαμε το σύστημα που αναπτύχθηκε για τον εντοπισμό των όρων που αναφέρονται σε κείμενα που προέρχονται από τρείς διαφορετικές πηγές δεδομένων. Τους όρους αυτούς, τους συγκρίνουμε με ιατρικούς όρους που έχουν ενδιαφέρον για το Σύνδρομο Σιόγκρεν και οι οποίοι είχαν καθοριστεί εξαρχής από ειδικούς στο χώρο της ιατρικής έρευνας με την μορφή μιας οντολογίας, στα πλαίσια του έργου HarmonicSS [4]. Τέλος, εξάγουμε συμπεράσματα επάνω στα αποτελέσματα μας για τον αλγόριθμο C-Value, προτείνουμε εναλλακτικές διαφοροποιήσεις που μπορούν να γίνουν καθώς και επιπρόσθετες τεχνολογίες που μπορούν να ενισχύσουν τα αποτελέσματα μας.

1.2 Δομή Εργασίας

Στην ενότητα (2) ασχολούμαστε με περεταίρω ανάλυση των μεθόδων που σχετίζονται με τον Εντοπισμό Όρων, με εστίαση σε μία διαδεδομένη τεχνική εξαγωγής Όρων πολλαπλών λέξεων η οποία ονομάζεται C-Value και ασχολούμαστε με τη γενικότερη βιβλιογραφία η οποία έχει δημιουργηθεί για την αναγνώριση όρων. Γίνεται μία περίληψη των γενικότερων τάσεων που επικρατούν γύρω από τις μεθόδους εντοπισμού τεχνικών όρων, ενώ αναφέρονται και διάφορες μετρικές που έχουν δημιουργηθεί μέσα από πειράματα από ερευνητικές ομάδες ώστε να ποσοτικοποιηθεί ο βαθμός στον οποίο μία λέξη ή μια φράση αποτελεί έναν όρο.

Στη συνέχεια, στην ενότητα (3) περιγράφουμε την προσέγγιση που ακολουθήθηκε η οποία βασίστηκε στον αλγόριθμο υπολογισμού του CValue, προκειμένου να αναπτύξουμε την εφαρμογή μας, τον επονομαζόμενο *TermFinder*. Εξηγούμε πώς με την εφαρμογή προσεγγίζουμε τον στόχο που έχουμε να υποβοηθήσουμε την ιατρική προσπάθεια γύρω από το σύνδρομο Σιόγκρεν και τις ανάγκες που προσπαθούμε να καλύψουμε με αυτή την εφαρμογή, ενώ στις ενότητες (0) και (3.3) μπαίνουμε σε μεγαλύτερο βάθος στο τεχνικό σκέλος του προγράμματος με λεπτομέρειες σε αρχιτεκτονικό και λειτουργικό επίπεδο.

Έπειτα, ασχολούμαστε με τις εκτελέσεις του αλγορίθμου για αναγνώριση ιατρικών όρων που παραπέμπουν σε θεραπείες, εξετάσεις και άλλες σχετικές πληροφορίες με το σύνδρομο Σιόγκρεν.

Η ενότητα (0) περιλαμβάνει τις δοκιμαστικές εκτελέσεις του προγράμματος μας σε γνωστά inputs από τα οποία γνωρίζουμε τι output περιμένουμε επακριβώς. Στα σύντομα αυτά δοκιμαστικά inputs, εντοπίζουμε αρχικά από μόνοι μας τις εκφράσεις που θα περιμέναμε να είναι όροι (σε ιδανικές συνθήκες, αυτό θα έπρεπε να εκτελεστεί από ιατρούς-ειδικούς, αλλά στα πλαίσια αυτής της

διπλωματικής αυτό δεν έγινε) και μετά τα εισάγουμε στον αλγόριθμο μας. Με βάση τα αποτελέσματα που βγάζουμε αξιολογούμε το πρόγραμμα μας εισάγοντας τις εξόδους του στον Αλγόριθμο Αξιολόγησης Αποτελεσμάτων.

Η ενότητα (4.2) περιέχει ανάλυση επί της μεθοδολογίας που ακολουθήσαμε για τα πειράματα, με αρχική επεξήγηση των datasets εισόδου, επεξήγηση των διαμορφώσεων της εφαρμογής και των παραμέτρων που εξετάσαμε καθώς και μία ανάλυση επί των αποτελεσμάτων, με παρατηρήσεις επί των όρων που ανιχνεύθηκαν, αγνοήθηκαν αλλά και βρέθηκαν με την τεχνική του C-Value.

Η σύνοψη στην ενότητα (0) αναδεικνύει όσα αποκομίσαμε από ολόκληρη την εργασία και τα συμπεράσματα που εξήχθησαν. Εμπεριέχει κατά κάποιον τρόπο τις σημαντικότερες παρατηρήσεις από τα αποτελέσματα αλλά και έναν σύντομο συνολικό σχολιασμό πάνω στην τεχνική του C-Value, ενώ επιχειρούμε να προτείνουμε και πιθανές βελτιώσεις που θα μπορούσαν να γίνουν επί του αλγορίθμου αλλά και κατευθύνσεις που κρίνουμε ότι έχουν νόημα να διερευνηθούν περεταίρω.

Στο παράρτημα (7) παραθέτουμε την αρχιτεκτονική του Αλγορίθμου Δοκιμής και Αξιολόγησης Συστημάτων, που χρησιμοποιήθηκε συμπληρωματικά για την διεξαγωγή των Πειραμάτων αλλά και για την μέτρηση της απόδοσης του *TermFinder*. Επιπλέον, εμπεριέχονται ανεπεξέργαστα δεδομένα από τις εξόδους του *TermFinder* επί των πειραμάτων στην ενότητα (0), τα οποία τα συστήνουμε για ανάγνωση κατά τη διάρκεια μελέτης της προαναφερθείσας ενότητας για καλύτερη κατανόηση των παρουσιαζόμενων πινάκων και προτεινόμενων όρων.

2 State of the Art

2.1 Επεξεργασία Κειμένου – Εντοπισμός Όρων

Η έρευνα μας ξεκινά με την υπόθεση ότι το πρόβλημα της Αυτόματης Αναγνώρισης Όρων μπορεί να αντιμετωπιστεί αποδοτικά υπό το πρίσμα της Επιστήμης της Επεξεργασίας κειμένου, με αξιοποίηση ήδη ανεπτυγμένων αλγορίθμων της. Αξία έχει να γίνει αρχικά μία σύντομη ιστορική αναφορά στον κλάδο της Επεξεργασίας Κειμένου, με αναφορά στους μηχανισμούς που έχουν κατά καιρούς χρησιμοποιηθεί αλλά και στις θεωρητικές τάσεις που εξελίσσονταν.

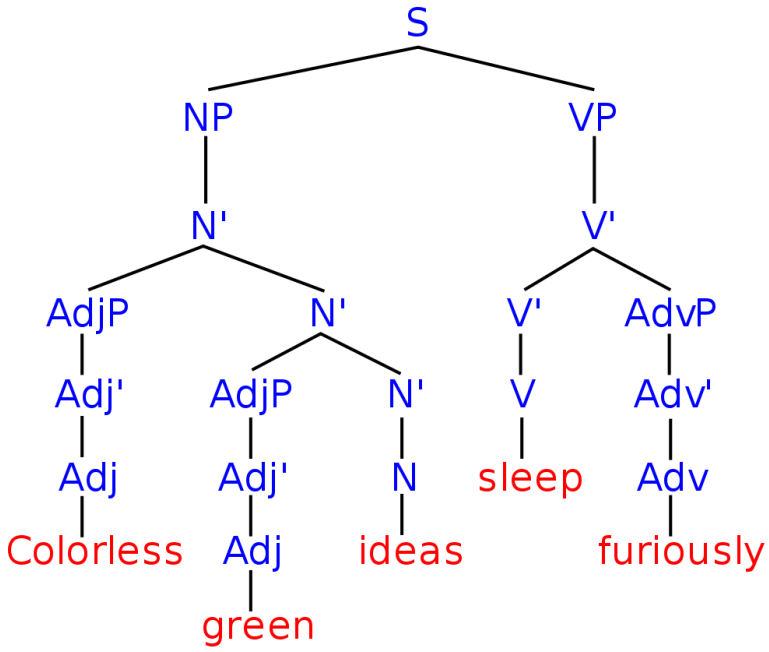
Η τεχνολογική επανάσταση στην Πληροφορική κατά τα μέσα του 20ού αιώνα οδήγησε στην συνεχιζόμενη ενίσχυση της διαθέσιμης υπολογιστικής ισχύος, την οποία προσπάθησαν να αξιοποιήσουν νεοεμφανιζόμενα παρακλάδια από άλλες επιστήμες, ώστε να τιθασεύσουν αυτή τη νέα ισχυρή δύναμη. Η εμφάνιση του κλάδου της Υπολογιστικής Γλωσσολογίας ήρθε σε αυτό το πλαίσιο στην Αμερική κατά τη δεκαετία του ‘50, στην αρχή με μικρό θεωρητικό υπόβαθρο κυρίως με τον πρακτικό στόχο να αντιμετωπιστεί το πρόβλημα της Μηχανικής Μετάφρασης. Υπήρχε η πεποίθηση, αρχικά, ότι η μεγάλη ποσότητα κειμένων από διεθνή βιβλιογραφία μπορεί με ευκολία να μεταφραστεί στην Αγγλική με αυτοματοποιημένο αλγοριθμικό τρόπο, όμως ήδη από τα πρώτα χρόνια ερευνών, διαπιστώθηκε η πολυπλοκότητα του έργου. Για αυτό, οι προσπάθειες οργανώθηκαν σε μια νεοσύστατη επιστήμη, την *Υπολογιστική Γλωσσολογία,* «την μελέτη των υπολογιστικών συστημάτων για την κατανόηση και τη δημιουργία φυσικής γλώσσας». [5]

Η Υπολογιστική Γλωσσολογία ξεπέρασε τις πρώτες άκαρπες προσπάθειες Μηχανικής Μετάφρασης και ακολουθεί ένα θεωρητικά-βασισμένο μοντέλο, στηριζόμενο στη Θεωρία Αυτομάτων, με στόχο να δομήσει αυστηρούς κανόνες γραμματικής και να προβλέπει επακριβώς τη δημιουργία των οντοτήτων μίας γλώσσας [6]. Τα επόμενα χρόνια ακολουθεί πρόοδος στον τομέα με τη δημιουργία του Chomsky Ierarchy, context-free grammars και εφαρμογές σε άλλους τομείς, όπως στη δημιουργία συντακτικού προγραμματιστικών γλωσσών, όπως compilers.

Ένας από τους βασικούς πρωτοπόρους της επιστήμης, ο Noam Chomsky, δήλωσε το 1956 ότι «no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar» [7]. Συνοπτικά μιλώντας, το σύνολο των κανόνων γραμματικής δεν καλύπτει τις ανάγκες της σημασιολογίας για μία πραγματική γλώσσα, καθώς πέρα από το να ερμηνεύει τη σύνταξη

του λόγου σωστά, δεν κατανοεί τις «εξαιρέσεις» ενός πιο ανεπίσημου και αδόμητου λόγου στη γραμματική, αλλά ούτε μπορεί να αντιληφθεί τις υποκείμενες σημασιολογίες που είναι απαραίτητο συστατικό μίας γλώσσας.

Σε αυτήν την διαπίστωση ήρθε να αναφερθεί και το περίφημο παράδειγμα του Chomsky στην αγγλική γλώσσα: "*Colorless green ideas sleep furiously*"[1]. Μια φράση πλήρως σωστή γραμματικά και γεννημένη από ένα Chomsky Normal Form (CNF), μία context-free γραμματική, μπορεί να στερείται πλήρως του νοήματος που αναζητούμε, και να μην κρύβει καν κάποιον πραγματικό όρο, σε περίπτωση που θέλουμε να εξάγουμε από αυτήν κάποιο νόημα με δομημένη διαδικασία.



Σχήμα 1: Δομή της συντακτικά σωστής, αλλά χωρίς σημασία, Αγγλικής πρότασης "Colorless green ideas sleep furiously", Chomsky 1957

Έτσι, για να αντιμετωπιστεί η αδυναμία της γραμματικής και του συντακτικού από μόνες τους να αντιμετωπίσουν με δομημένο τρόπο προβλήματα σχετικά με την κωδικοποίηση της γλώσσας χρειάστηκε να αναπτυχθούν οι επονομαζόμενες *στατιστικές μέθοδοι επεξεργασίας κειμένου.*

Αυτές οι μέθοδοι αποτελούνται από εργασίες τις οποίες επιτελεί ένα σύστημα επεξεργασίας κειμένου τις οποίες διαχωρίζουμε σε *χαμηλού* (low level) και *υψηλού* (high level) επιπέδου [8]. Αυτή η κατηγοριοποίηση προκύπτει ανάλογα με το πόσο συγκεκριμένη και απαραίτητη είναι η δουλειά για την βασική περαίωση της Αυτόματης Εξαγωγής Όρων. Ιδιαίτερη σημασία πρέπει να δοθεί στις πρώτες,

[1] Ιστορική φράση του Chomsky το 1956 για να δείξει ότι το ορθό συντακτικό δεν βγάζει πάντα νόημα https://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously

όπως tokenization, POS tagging, spelling error identification and recover, named entity recognition μιας και οι υψηλού επιπέδου αφορούν και εργασίες μετά την εύρεση των όρων, εξειδικευμένες σε συγκεκριμένα προβλήματα. Ειδικά το τελευταίο, προσομοιάζει με το term recognition και χρειάζεται να διερευνηθούν η σειρά των λέξεων, η διαφοροποίηση του μέρους του λόγου τους, αλλαγές στις πτώσεις τους, συνωνυμίες αλλά και ομοηχίες με μη σχετικές λέξεις. Παράλληλα, χρειάζεται και ανάλυση στα συμφραζόμενα τους, για τα οποία πολλές φορές είναι χρήσιμη και η εξαγωγή σχέσεων μεταξύ τους είτε σε επίπεδο λέξεων είτε ακόμα και σε επίπεδο εννοιών, στην οριζόντια αλλά και στην κατακόρυφη κλίμακα (δηλαδή ακόμα και στο αν είναι υποσύνολα/υπερσύνολα μεταξύ τους).

Η επεξεργασία φυσικής γλώσσας έρχεται να δώσει απάντηση στον μεγάλο όγκο δεδομένων που παράγεται στον ιατρικό κλάδο υπό μορφή κειμένων και άρθρων, μιας και παρέχει την απαραίτητη δυνατότητα πρόσβασης στην πληροφορία που κρύβουν μέσα από την ταυτοποίηση όρων. Η εξαγωγή της ουσιαστικής και δομημένης πληροφορίας των άρθρων από τα υπολογιστικά συστήματα επιλύει το πρόβλημα επεξεργασίας μεγάλων δεδομένων στον ιατρικό τομέα.

Η προσέγγιση του θέματος ως απλή δημιουργία και συντήρηση ενός λεξικού (σημασιολογική προσέγγιση) υστερεί λόγω της διαρκούς μεταβολής και δημιουργίας των όρων, της διαφορετικής ερμηνείας τους ή και της πολλαπλής ονομασίας τους, θέματα τα οποία μια αυτοματοποιημένη διαδικασία ταυτοποίησης όρων μπορεί να προσπεράσει. Αυτή η διαδικασία θα ακολουθεί τα εξής τρία βήματα: αναγνώριση όρων, ταξινόμηση όρων και καθορισμός (όρων) [9]. Το πρώτο βήμα αφορά τον εντοπισμό στο κείμενο λέξεων ή φράσεων που υποδηλώνουν την παρουσία όρων σχετικών με τον τομέα στον οποίο αναφέρεται το κείμενο. Το δεύτερο αφορά την κατηγοριοποίηση των όρων σε επιμέρους κλάδους του τομέα που μελετούμε, την περεταίρω δηλαδή διερεύνηση του ρόλου που κατέχουν μέσα στον τομέα. Τέλος, ο καθορισμός ή χαρτογράφηση (mapping) του όρου συνδέει τον όρο που ανιχνεύσαμε με καλά ορισμένες έννοιες του επιστητού και επισφραγίζει την ταυτότητα του. Τυχούσα υποβοήθηση στη συσχέτιση των όρων με τη βιβλιογραφία μπορεί να λάβει χώρα μέσα από μία συνιστώσα κανονικοποίησης ήδη από το στάδιο αναγνώρισης των όρων.

Η επιτυχία της αναγνώρισης όρων βασίζεται σε 2 μετρικές, στην *ακρίβεια (precision)* που αφορά την ορθότητα της εύρεσης των όρων και στην *ανάκληση (recall)* που αφορά το ποσοστό του πλήθους των ορθών όρων που εντοπίσαμε επί του συνόλου των πραγματικών αυτών [9]. Η σχέση των 2 μετρικών είναι κατά βάση αντιστρόφως ανάλογη – αν αυξήσουμε τους όρους που μαντεύουμε, η ανάκληση θα αυξηθεί αφού αυξάνουμε τις πιθανότητες να βρούμε κάποιον πραγματικό, όμως η ακρίβεια μειώνεται αφού έχουμε και πολλά false positives. Μπορούμε να χρησιμοποιήσουμε έναν πιο σύνθετο τύπο της *F-measure* που εξαρτάται μόνο από τις 2 άνω μετρικές για να έχουμε συνολική εικόνα.

Γενικότερα, οι βασικότερες προσεγγίσεις στην αναγνώριση όρων διαχωρίζονται σύμφωνα με το αν βασίζονται σε έτοιμα λεξικά για την -ουσιαστικά- αναζήτηση όρων, σε κανόνες-μοτίβα τα οποία ανιχνεύουν σχηματισμούς (συντακτικούς ή και σημασιολογικούς) υποψήφιων όρων και τέλος σε τεχνικές στατιστικής και μηχανικής μάθησης. Δεν αποκλείονται επιπλέον και οι υβριδικές προσεγγίσεις με συνδυασμό κάποιων εκ των προηγούμενων ώστε να εξασφαλιστεί η μεγαλύτερη δυνατή F-measure. Συνοπτικά, παρατηρείται ότι οι στατιστικές μέθοδοι εξασφαλίζουν τα καλύτερα συγκριτικά αποτελέσματα, ιδιαίτερα όταν εντάσσονται σε αυτές τμήματα από τις υπόλοιπες μεθοδολογίες (π.χ. C/NC-value). Αναφορά αξίζει επίσης και στον κλάδο της αναγνώρισης ακρωνυμίων και εύρεσης των «αναπτυγμένων μορφών» τους (expanded forms).

Όσον αφορά το 2ο βήμα, δηλαδή την κατηγοριοποίηση όρων, παρατηρούμε μία ανάποδη τάση από αυτήν της αναγνώρισης όρων. Πιο συγκεκριμένα, εδώ οι μέθοδοι που βασίζονται σε εξωγενή τροφοδότηση πληροφορίας (π.χ. με χρήση category-specific word lists και πληροφορίας σχετικής με επιθυμητά Part-Of-Speech) πετυχαίνουν καλύτερα αποτελέσματα από αυτά των ενδογενών μεθόδων [9] (π.χ. στατιστική προσέγγιση με Bayesian Method). Όμως και εδώ, οι υβριδικές μέθοδοι που αξιοποιούν και τις 2 προσεγγίσεις (π.χ. ομοιότητα όρων εσωτερικά αλλά και με εξωτερικά τροφοδοτημένες λίστες όρων) οδηγούν σε μεγαλύτερη επιτυχία.

Τέλος, για το 3ο βήμα, την χαρτογράφηση δηλαδή των όρων σε ένα ήδη υπάρχον σχήμα αναφοράς με ίδια κλειδιά, καλούμαστε να αντιμετωπίσουμε τα προβλήματα της αντιστοίχισης n-σε-1 όρο όπως και την 1-σε-m όρους. Διαφορετικές στρατηγικές που μπορούμε να ακολουθήσουμε είναι για παράδειγμα το stemming και η επακόλουθη ορολογική αναζήτηση σε εξωτερικά λεξικά για συνώνυμα, τεχνικές μηχανικής μάθησης/κανόνων για αξιολόγηση συνωνύμων με βάση τα συμφραζόμενα ή κάποια γνωστά μοτίβα.

Σημαντικό είναι να επιχειρηθεί μια αυστηρότερη καταγραφή των διαφόρων εννοιών που απαρτίζουν την γενικότερη Αυτόματη Εξαγωγή Όρων και συνιστούν διεργασίες που την χαρακτηρίζουν. Στόχος της είναι η *«αναγνώριση του βασικού λεξιλογίου ενός εξειδικευμένου τομέα»* και *«νοείται πρώτα και κύρια ως η εξ-υπολογιστών βοήθεια για ανακούφιση αυτής της χρονοβόρας εργασίας»* [10]. Βασικό είναι να τονιστεί ότι η ολοκληρωμένη σημασιολογική μοντελοποίηση των όρων είναι αδύνατη από τους υπολογιστές για αυτό και ο ρόλος τους παραμένει υποβοηθητικός.

Οι εργασίες της Αυτόματης Εξαγωγής Όρων μπορούν να εντοπιστούν στις εξής, τη συλλογή των όρων (*Corpus Collection*), τον εντοπισμό της *Unithood,* τον εντοπισμό της *Termhood*, τον εντοπισμό των παραλλαγών όρων *(Term Variants)*, την αξιολόγηση και την επαλήθευση. Στη συνέχεια, η ίδια η

Αυτόματη Εξαγωγή Όρων ακολουθείται από εργασίες όπως η Διαχείριση Ορολογιών, η Υποστήριξη Μεταγλώττισης και η Εξαγωγή Πληροφορίας, οι οποίες επωφελούνται από τα αποτελέσματα της.

Για τη συλλογή των κειμένων, χρησιμοποιούνται δειγματοληπτικές τεχνικές, καθώς επίσης έχουν προταθεί και συμπληρωματικές διεργασίες όπως οι χρήση επονομαζόμενων "seed terms" [10] δηλαδή επιλεγμένων από ανθρώπους ως input χαρακτηριστικών όρων για να συλλεχθούν σχετικά με τον κλάδο κείμενα, μέσα από την αξιολόγηση της σχετικότητας των περιεχομένων τους με τους άνωθι όρους. Αυτοί, ανατροφοδοτούνται στα επόμενα βήματα αυτόματα από την αυτοματοποιημένη διαδικασία και χρησιμοποιούνται οι προηγούμενες έξοδοι ως νέοι seed terms.

Σχετικά με το unithood, δηλαδή *«τον βαθμό της ισχύος ή σταθερότητας των συντακτικών συνδυασμών και συγκατατάξεων»* [10] σημειώνεται ότι είναι σημαντικός καθώς οι περισσότεροι όροι είναι φράσεις και όχι λέξεις, δεύτερον γιατί οι ίδιοι οι όροι αντιστοιχούν 1 προς 1 με την έννοια που αναπαριστούν και τρίτον γιατί στην αρχή θεωρούταν ταυτόσημος με το termhood καθώς οι συνδυασμοί των λέξεων ταυτίζονταν με το αν ήταν όρος ή όχι αυτός ο σχηματισμός. (Στην συνέχεια αποδεικνύεται ότι αυτός ο ισχυρισμός αποδεικνύεται μη αληθής)

Σχετικά με το termhood, δηλαδή «τον βαθμό στον οποίο μια σταθερή λεκτική οντότητα σχετίζεται με κάποιες εξειδικευμένες (με τον τομέα μελέτης) έννοιες» σημειώνεται ότι αποτελεί αρκετές φορές διαφορετική μετρική από το unithood, δεδομένου ότι συχνές αλλά καθημερινές εκφράσεις έχουν υψηλό unithood γιατί εμφανίζονται με την ίδια δομή πολλές φορές αλλά έχουν χαμηλό termhood γιατί δε σχετίζονται με κάποια σχετική με τον τομέα έννοια. Όμοια, ασυνήθιστοι και μονολεκτικοί ορισμοί έχουν ψηλό termhood αλλά χαμηλό unithood. Για τον υπολογισμό του termhood, αρχικά χρησιμοποιούνταν διαφόρων ειδών συχνότητες, όπως η απόλυτη συχνότητα, η σχετική συχνότητα δια μέσου πολλών διαφορετικών αρχείων, η σύγκριση των συμφραζομένων ή και η ίδια η μορφολογική ανάλυση των επιμέρους λέξεων και συνθετικών.

Η αντιπαραβολική προσέγγιση στην εξαγωγή όρων *(Contrastive Term Extraction)* [10] είναι μία θεώρηση ότι οι όροι εμφανίζονται συχνά σε κείμενα του τομέα τους και σπάνια σε κείμενα γενικότερης καθημερινής χρήσης. Αυτή ουσιαστικά οδηγεί σε μία νέα μετρική, την *weirdness* μιας λέξης η οποία φαίνεται να σχετίζεται με την πιθανότητα να αποτελεί όρο.

2.2 Αυτόματος Εντοπισμός Όρων

Παρατηρώντας την αναγκαιότητα αυτοματισμού στην δημιουργία ορολογιών λόγω του μεγάλου εισερχόμενου όγκου νέων τεχνικών όρων, διαπιστώνει κανείς ότι η προσπάθεια αυτή μπορεί να ακολουθήσει την λογική 2 ήδη υπαρχόντων και ξεχωριστών κλάδων, αυτός των «*Κύκλων Ανάκτησης*

Πληροφορίας» (Information Retrieval Circles) καθώς και της «*Υπολογιστικής Γλωσσολογίας» (Computational Linguistics)* [11].

Στον τομέα της Ανάκτησης Πληροφορίας, στόχος είναι η αντιστοίχιση πληροφορίας και αρχείων με αιτήματα, και συχνά αξιοποιούνται τεχνικές *Αυτόματης Ευρετηρίασης (Automatic Indexing)*, με βάση τις οποίες εντοπίζονται αντιπρόσωποι του περιεχομένου τους και χρησιμοποιούνται αυτοί για την αντιστοίχιση. Οι αντιπρόσωποι είναι λεκτικές μονάδες ή οντότητες, «λέξεις» και *Όροι Ευρετηρίασης (Index Term)*, που επιλέγονται ανταγωνιστικά μεταξύ τους με βάση κάποιου είδους στάθμισης και επαληθεύονται με βάση *την ανάκληση και την ακρίβεια τους (Recall and Precision).*

Μεγάλη ανάλυση έχει γίνει στις μεθόδους διαφορετικής *στάθμισης (weight)* των «όρων», ανάλογα με τις λέξεις που τους αποτελούν και τα κείμενα στα οποία εντοπίζονται. Κάποιες από αυτές στηρίζονται στις απόλυτες συχνότητες εμφάνισης σε επίπεδο αρχείου, βάσης και *δια-αρχειακής κατανομής (Cross-Document Distribution)* [11]. Η κατανομή αυτή υπολογίζεται εις διπλούν, ως Poisson στα σχετικά και μια δεύτερη όμοια στα μη σχετικά κείμενα. Όπως και να έχει, υπάρχει σημαντική συσχέτιση ανάμεσα στους Όρους Ευρετηρίασης με τους Τεχνικούς Όρους, ώστε οι τεχνικές υπολογισμού του πρώτου να υπαχθούν στο δεύτερο.

Αναγκαίος είναι και ο διαχωρισμός των μετρικών *unithood* και *termhood*, καθώς αποτελούν διαφορετικές προσεγγίσεις εξ'ολοκλήρου. Το Unithood αφορά των βαθμό ισχύος ή σταθερότητας των συνδυασμών λέξεων σε συντακτικό επίπεδο, ενώ το termhood αφορά τον βαθμό που μία γλωσσική μονάδα συνδέεται σε εννοιολογικό επίπεδο με τον τομέα μελέτης. Συνοπτικά, στην στατιστική προσέγγιση του προβλήματος της Αυτόματης Αναγνώρισης Όρων, υπάρχουν οι τεχνικές των n-grams (log-likelihood), Structural Dependency, C-Value και άλλες, που υπάγονται ξεχωριστά η καθεμία στην εξαγωγή κάποιας από τις 2 προαναφερθείσες μετρικές. Η τεχνική του log-likelihood αφορά το κατά πόσο συσχετίζεται μία ακολουθία n-λέξεων (n-grams) με ένα προτεινόμενο μοντέλο και κατανομή εμφανίσεως όρων, ενώ με την C-Value θα ασχοληθούμε εκτενέστερα σε επόμενο κεφάλαιο, μιας και θα αποτελέσει το κεντρικό συστατικό για το εργαλείο που θα αναπτύξουμε.

Όσον αφορά το *unithood*, ένα unit το οποίο εμφανίζεται συχνά, αποκλειστικά, κυρίως ή με κάποιο μοτίβο σε έναν τομέα είναι πιθανό να είναι όρος. Όσον αφορά το *termhood*, ένας σύνθετος όρος μπορεί είτε να σχετίζεται εν μέρει, είτε να μην σχετίζεται με τα απαρτιζόμενα μέρη, ή να εξαρτάται 1 προς 1 με την σύνδεση τους, δηλαδή σε αυτή την περίπτωση, το termhood ταυτίζεται με το unithood. Σε κάθε περίπτωση, η πειραματική -δεδομένου ότι η θεωρητική είναι δύσκολη- θεμελίωση των τεχνικών γίνεται ως προς τη φύση των δεδομένων που μελετούμε αλλά και ως προς τα χαρακτηριστικά που αναμένουμε να έχουν οι παραγόμενοι όροι.

Στο paper Automatic Term Recognition based on Statistics of Compound Nouns and their Components [12] αναλύεται η συσχέτιση ανάμεσα σε όρους που είναι σύνθετα ουσιαστικά με τα επιμέρους συνθετικά τους, όταν κι αυτά αποτελούν υποψήφιους όρους. Η δομική μελέτη των όρων αποκτά ιδιαίτερη σημασία από τη στιγμή που η πλειοψηφία εξ 'αυτών συνιστούν σύνθετα ουσιαστικά, τα οποία απαρτίζονται από μικρότερα ουσιαστικά. Επομένως, η συνεισφορά της συσχέτισης των επιμέρους συνθετικών (δομή) ενός υποψήφιου όρου στην *termhood (βαθμός που μία λεκτική οντότητα σχετίζεται εννοιολογικά με έναν τομέα)* είναι σημαντική, αφού μεταφράζεται μία προς μία με την συσχέτιση των επιμέρους εννοιών σε έναν τομέα. Για παράδειγμα, ο συνδυασμός της λέξης πόνου ως σύμπτωμα στον ιατρικό τομέα και κεφάλι ως μέρος του σώματος, αναδεικνύει την δομική εξάρτηση που έχει γενικά ένα σύμπτωμα με ένα μέρος του σώματος (και άρα αποκτά νόημα ο όρος πονοκέφαλος). Όμοια, μπορούμε να δουλέψουμε αντίστροφα για να βρούμε από το unithood των μερών μιας έκφρασης το termhood του συνόλου. Προς αυτή την κατεύθυνση, απαιτείται μια διαφορετική μελέτη, που αφορά τον *εννοιολογικό χώρο (term space) ενός τομέα (domain)* για την οποία οι συγγραφείς θεωρούν ότι χρειάζεται περισσότερη έρευνα. Αρκεί να ειπωθεί ότι ο εννοιολογικός χώρος αφορά τους κανόνες με τους οποίους κατασκευάζονται όροι με συναρμολόγηση των ήδη υπαρχουσών εννοιών ενός τομέα, και πιθανή κατανόηση του οποίου θα οδηγούσε αυτόματα σε ακριβέστερο προσδιορισμό των πραγματικών όρων.

Τα στατιστικά ευρήματα στο πλαίσιο του εννοιολογικού χώρου εκφράζουν σε μεγαλύτερο βαθμό το «νόημα του συγγραφέα», από τη στιγμή που ισχύει η παραδοχή ότι ένας όρος που θέλει να επισημάνει ο συγγραφέας ως σημαντικό στον κείμενο, όχι μόνο θα εμφανίζεται πολλές φορές αλλά και με πολλαπλές μορφές, πολλές φορές σύνθετες λέξεις και μάλιστα, νεολογισμούς (δηλαδή νέες δημιουργημένες λέξεις). Επομένως, ο εννοιολογικός χώρος ως σύνολο των σχετιζόμενων συνθετικών/υπο-όρων μίας λέξης πρέπει να λαμβάνεται υπόψη στον υπολογισμό του termhood.

Η υπόθεση που κάνουν οι συγγραφείς Nakagawa H. και Mori T. είναι ότι η σύνθετη δομή σε όρους προκύπτει από άλλους ήδη υπάρχοντες απλούστερους όρους. Αυτή, λαμβάνεται στην πράξη υπόψιν στον υπολογισμό που έχουν προτείνει άλλοι ερευνητές (Frantzi-Ananiadou 1996 [13]) με την μετρική C-Value και NC-Value. Σημασία έχει και το πλήθος αυτών των υποόρων, με την πλειοψηφία να αποτελούνται από 2 (bigrams) και άλλα από 3 (trigrams). Άλλη μετρική που είναι σημαντική είναι το αν αυτοί οι υποόροι βρίσκονται δεξιά ή αριστερά του βασικού όρου *(#LN, #RN)* και σε τί πλήθος. Επιπλέον, χρησιμοποιούμε έναν πολλαπλασιαστή για τον υποψήφιο όρο με βάση το πλήθος των φορών που εμφανίζεται «ανεξάρτητα» (δηλαδή χωρίς ουσιαστικά πλησίον του).

2.3 Τεχνική Αυτόματου Εντοπισμού Όρων (C-Value)

Η αυτοματοποιημένη επεξεργασία φυσικής γλώσσας από έναν υπολογιστή αποτελεί έναν ερευνητικό τομέα της *Υπολογιστικής Γλωσσολογίας*, ένα αντικείμενο της οποίας είναι και η αναγνώριση τεχνικών όρων από ψηφιακές βιβλιοθήκες. Ιδιαίτερη αξία έχει, το αυτοματοποιημένο σύστημα να εξάγει τους όρους αυτούς, χωρίς να τροφοδοτηθεί εξωγενώς με πληροφορία σχετικά με τον τεχνικό τομέα τον οποίον αφορούν, δεδομένου ότι η εξέλιξη του τομέα αυτού συνεπάγεται διαρκή δημιουργία νέων όρων. Οι μέθοδοι C-Value και NC-Value [13] χρησιμοποιούνται για τον σκοπό αυτόν ακριβώς, μιας και πρόκειται για αλγορίθμους που αξιοποιούν τόσο γλωσσολογική όσο και στατιστική ανάλυση του κειμένου προκειμένου να εντοπίσουν τους σημαντικότερους από αυτούς τους όρους.

Πιο συγκεκριμένα, στην ανάλυση μας στους όρους αυτούς, ξεπερνάμε τον περιορισμό της μοναδικής λέξης και περνάμε στην ουσία στην ανάλυση φράσεων-όρων με πολλαπλές εμφωλεύσεις. Σε αυτές τις εμφωλεύσεις, χρησιμοποιούμε τον αλγόριθμο C-Value για να υπολογίσουμε το termhood. Η λειτουργία του είναι να ανιχνεύει το termhood δηλαδή «την στατιστική απεικόνιση των υποψήφιων όρων» [13] που το σύστημα μας έχει υπολογίσει για αυτούς, ο καθένας εκ τους οποίους εντοπίζεται μετά από την εφαρμογή διαδοχικών γραμματικών και συντακτικών φίλτρων.

Στο πρώτο στάδιο, που αφορά την γλωσσική επεξεργασία του κειμένου, έχουμε την συστηματική γραμματική και συντακτική επεξεργασία των γραμματοσειρών του κειμένου. Ξεκινάμε με την αναγνώριση του μέρους του λόγου για κάθε γραμματοσειρά έτσι ώστε στη συνέχεια να φιλτράρουμε ως προς τα μέρη και την αλληλουχία τους που μας ενδιαφέρουν με την αξιοποίηση κανονικών εκφράσεων. Στη συνέχεια, συμβουλευόμαστε μία «stoplist» ή λίστα εξαιρέσεων που απαρτίζεται από συχνές λέξεις οι οποίες δεν αποτελούν επιβεβαιωμένα όρους.

Στο δεύτερο στάδιο, έχουμε την στατιστική ανάλυση των υποψήφιων όρων που συνοψίζεται στη διαδικασία υπολογισμού του C-Value/termhood η οποία είναι μια τιμή που λαμβάνει υπόψη θετικά τη συχνότητα εμφάνισης της υποψήφιας συμβολοσειράς τόσο απόλυτα όσο και σχετικά με τις άλλες καθώς επίσης και το μάκρος τους. Αρνητικά συνυπολογίζεται το πόσο συχνά αυτές οι συμβολοσειρές αποτελούν υποσύνολο μεγαλύτερων υποψήφιων όρων. Η ανάλυση γίνεται επαναληπτικά κατά μειούμενο πλήθος λέξεων.

Τέλος, υπολογίζουμε μια δεύτερη τιμή, την NC-Value η οποία συμπεριλαμβάνει την επιρροή από τα συμφραζόμενα των όρων. Πιο συγκεκριμένα, πρόκειται στην ουσία για έναν επανυπολογισμό του termhood αναθέτοντας βάρη στις λέξεις (ουσιαστικά επίθετα και ρήματα) των όρων σύμφωνα με τον πρώτο υπολογισμό της C-Value για κάθε όρο αλλά και το άθροισμα του πλήθους τους. Όροι με ψηλό

C-Value αποκτούν μεγαλύτερο βάρος και τελικά τα βάρη αυτά συναθροίζονται με το κλασσικό C-Value. Έτσι, εξασφαλίζουμε μεγαλύτερη ακρίβεια στους πιο συχνούς (και πιθανά σημαντικούς όρους), μειώνοντας την ακρίβεια στους λιγότερο σημαντικούς. Η μέθοδος αυτή, διατηρεί τον αυτοματοποιημένο χαρακτήρα και την ανεξαρτησία του συστήματος καθώς δεν αξιοποιεί εξωγενή φίλτρα.

Αξιοσημείωτο είναι το γεγονός ότι πολλές φορές αυξάνουμε την ακρίβεια μας εντάσσοντας ως «συμφραζόμενους όρους» και ΜΗ-όρους, δεδομένου ότι πολλές φορές εμφανίζονται κοντά σε όρους χωρίς να είναι οι ίδιοι όροι (π.χ. το ρήμα «ορίζω» δεν είναι όρος αλλά είναι συχνό συμφραζόμενο σε όρους).

Μία δεύτερη προσέγγιση στην αυτοματοποιημένη αναγνώριση όρων παρέχεται και από την εφαρμογή που αναπτύχθηκε στο πανεπιστήμιο του Cardiff στο πλαίσιο της έρευνας της δρ. Irena Spasić [14]. Το FlexiTerm, περί του οποίου γίνεται ο λόγος, βασίζεται σε μεγάλο βαθμό στις αρχές του αλγόριθμου για την παραγωγή του termhood (C-Value/NC-Value), μελετώντας την επίδραση στη χαλάρωση κάποιων αυστηρών κανόνων του στην ακρίβεια του αλγορίθμου.

Πιο συγκεκριμένα, όμοια με τις παραδοχές για το C-Value, τονίζεται ότι η αυτοματοποίηση της αναγνώρισης όρων είναι μεγάλης σημασίας. Οι ίδιοι οι όροι μπορεί να προέρχονται από πολλούς και διαφορετικούς τομείς για τους οποίους ο αναλυτής δεδομένων δεν διαθέτει ειδίκευση και θα ήταν αδύνατο να αντλήσει μόνος του τους σημαντικούς όρους ή θα υπήρχε σημαντικό κόστος για την αναζήτηση τους μαζί με έναν ειδικό. Με την παροχή τους αυτοματοποιημένα, ο οποίος μπορεί να λειτουργήσει, από τη μία, χωρίς την ανάγκη ενός εξωτερικού «λεξικού» και από την άλλη σε πληθώρα αρχείων εισόδου εξοικονομείται εξαιρετικός φόρτος εργασίας. Έτσι, η διαδικασία τροφοδότησης ενός υψηλότερου επιπέδου αναλυτή δεδομένων (αυτόματες περιλήψεις άρθρων, δημιουργία καταλόγων) καθίσταται βιώσιμη.

Η μεθοδολογία παροχής του C-Value έχει ήδη αναλυθεί, επομένως θα καλυφθούν οι διαφοροποιήσεις που ακολουθήθηκαν στην περίπτωση του FlexiTerm. Στην εφαρμογή αυτή, στο συντακτικό επίπεδο, υπάρχει μεγαλύτερη ανοχή στην αντιστοίχιση tokens από τον κλασσικό C-Value αλγόριθμο, με αξιοσημείωτη τη χαλάρωση στον κανόνα της σειράς στις λέξεις που απαρτίζουν έναν όρο. Στην περίπτωση του C-Term αντί να χρησιμοποιούνται ακολουθίες λέξεων, πρακτικά εφαρμόζονται σύνολα ή όπως ονομάζονται «bag-of-words» [14]. Η σειρά των λέξεων όπως μελετήθηκε οδηγεί λαθεμένα σε διαφοροποίηση όρων που κανονικά θα έπρεπε να εκλαμβάνονται ως ίδιοι.

Όμοια, στο γραμματικό επίπεδο, υπάρχει μεγαλύτερη ανοχή σε τυπογραφικά και ορθογραφικά λάθη, χρησιμοποιώντας προσεγγιστική αντιστοίχιση των όρων, βασιζόμενη μάλιστα στην φωνητική ομοιότητα τους (ώστε να προβλεφθούν πιθανές αβλεψίες κατά τη συγγραφή τους). Εδώ, έρχεται να βοηθήσει η μέθοδος του Edit Distance (ED) η οποία εφαρμόζεται και στο στάδιο παραγωγής των tokens. Το Flexiterm εφαρμόζει την μέθοδο αυτή με το Jazzy, ένα API ορθογραφίας.

Στο στατιστικό επίπεδο, χάρη στις προηγούμενες αλλαγές μπορούμε να αντιμετωπίζουμε τα candidate terms επί των οποίων εφαρμόζεται η C-Value ως σύνολα από tokens και όχι ακολουθίες, με αποτέλεσμα να είναι δυνατή η εφαρμογή πράξεων συνόλων στις περιπτώσεις που ψάχνουμε τους εμφωλευμένους όρους (κάτι εξαιρετικά σημαντικό για τη μέθοδο C-Value). Αυτό απλοποιεί σε μεγάλο βαθμό την πολυπλοκότητα του αλγορίθμου.

Πέρα από τις διαφοροποιήσεις από τον κλασσικό αλγόριθμο C-Value ιδιαίτερη μνεία χρήζει και στην διαδικασία ανάλυσης και παραγωγής συμπερασμάτων που ακολούθησε η ομάδα του Flexiterm. Οι έξοδοι του αλγορίθμου εμφανίζονται σε 3 μορφές, είτε ως πίνακες, είτε ως λίστες και τέλος είτε ως κανονικές εκφράσεις μοτίβα. Όλες εξ' αυτών μπορούν να χρησιμοποιηθούν για συγκρίσεις μεταξύ διαφορετικών διαδοχικών πειραμάτων ώστε να εξαχθούν χρήσιμα συμπεράσματα για την συχνότητα εμφάνισης των όρων μεταξύ διαφορετικών κατηγοριών κειμένων. Για παράδειγμα, στον κλάδο της ιατρικής, η ομάδα του Flexiterm σύγκρινε τις εξόδους της εφαρμογής όταν εφαρμοζόταν σε πολλαπλά datasets από τη βάση του PubMed σε αντιπαραβολή με μία λίστα από κλινικά κείμενα από νοσοκομεία. Όμοια, αντιπαραβολή μπορεί να γίνει και ως προς το ποιοι γράφουν τα άρθρα, αν είναι δηλαδή επιστήμονες του κλάδου Υγείας ή όχι. Τέλος, για την αξιολόγηση του dataset που παρήγαγε η εφαρμογή αυτοματοποιημένα, συγκρίθηκε με όμοιες εφαρμογές, οι οποίες έχουν επιβλεφθεί από τον Ιατρικό ακαδημαϊκό κλάδο.

Τα συμπεράσματα της έρευνας συνοψίζονται στο ότι η χαλάρωση του συντακτικού κριτηρίου (εφαρμογή bag-of-words) δεν έβλαψε την ακρίβεια διαχωρισμού όρων από μη όρους [14]. Επιπλέον, η ομαδοποίηση όρων βοηθά στην περίπτωση που στο dataset αναμένουμε να βρούμε λίγους όρους, ειδικά αν επεκταθεί με την αξιοποίηση σημασιολογικής ανάλυσης από κάποιο domain-free εξωγενές σύστημα. Καταληκτικά, επισημαίνεται η χρησιμότητα του αλγορίθμου στην μαζική επεξεργασία όγκου άρθρων για την παροχή γρήγορων συμπερασμάτων σε μεγάλους πληθυσμούς ασθενών, ταχύτερα από την παραδοσιακή μέθοδο των ερευνών με ερωταπαντήσεις.

Σε ένα άλλο paper [15], επεξηγείται μία απόπειρα εξαγωγής υποψήφιων όρων με 2 διαφορετικές μεθόδους, μία ευριστική μέθοδος με στατιστικούς υπολογισμούς (όπως μοτίβα ουσιαστικών, κανόνες εύρεσης όρων ανάμεσα σε εισαγωγικά και άλλα) και μία με τεχνικές μηχανικής μάθησης, σε κείμενα

γραμμένα σε γλώσσα Bengali-Hindi. Αναφέρεται από τους συγγραφείς η παρατήρηση ότι οι πολυπληθέστεροι όροι αποτελούνται από πολλαπλές λέξεις και συχνότερα με 2 *(bigram)*, καθώς επίσης και το γεγονός ότι είναι σημαντικός ο περεταίρω διαχωρισμός τους σε *ονομαστικά (bigram nominal)* (σ.σ. δηλαδή αποτελούμενο από ουσιαστικά) και μη.

Η *συνθετικότητα (compositionality)* των *εκφράσεων πολλαπλών λέξεων (MWE)*, δηλαδή ο «βαθμός στον οποίο τα χαρακτηριστικά των μερών μίας έκφρασης συνδυαζόμενα προβλέπουν τα χαρακτηριστικά ολόκληρης της έκφρασης» είναι μία μετρική που θέλουν να υπολογίσουν οι συγγραφείς και για αυτό καταφεύγουν στις μεθόδους εκτίμησης εκφράσεων που περιγράψαμε νωρίτερα, δηλαδή τις στατιστικές, τις υβριδικές και τις γλωσσολογικές. Παραλλαγές της συνθετικότητας είναι η συνολική συνθετικότητα, η μερική συνθετικότητα.

Η εστίαση έγκειται στην ανάπτυξη μιας προσέγγισης μηχανικής μάθησης που να αξιοποιεί την πληροφορία από πρότερες στατιστικές, συντακτικές και γλωσσολογικές μεθόδους. Έτσι, στον αλγόριθμο τους, μετά από το βήμα της προεπεξεργασίας, υπολογίζονται στατιστικές μετρήσεις όπως η log likelihood με είσοδο τις απόλυτες και τις σχετικές συχνότητες εμφάνισης για κάθε όρο, και η point-wise mutual information(pmi) με είσοδο διαδοχικές λέξεις και η οποία εκφράζει τον λόγο της πιθανότητας να βρεθούν μαζί προς τον λόγο της πιθανότητας συχνότητας να βρεθούν αλλού (π.χ. η λέξη Puerto με τη λεξη Rico έχουν πολύ υψηλή pmi(Puerto,Rico) αφού αποτελούν λέξεις που σπάνια συναντιόνται χώρια). Παράλληλα με αυτές τρέχει μια διεργασία stemming των λέξεων, ενώ αξιοποιείται ένα πακέτο σημασιολογικής συσχέτισης από το WordNet.

Τέλος, αφού αποκτήσουν και άλλες επιπρόσθετες γλωσσολογικές και συντακτικές μετρήσεις στο κείμενο τους, καταλήγουν σε ένα έτοιμο μείγμα δεδομένων για εκπαίδευση και δοκιμή μεθόδων μηχανικής μάθησης. Μια εξ'αυτών, η *Random Forest* χρησιμοποιείται γιατί παράγει έναν ακριβή *ταξινομητή (classifier)*, τρέχει σε μεγάλο όγκο features (C-Value κ.ο.κ) και εκτιμά σωστά τα περιθώρια λάθους, με σύγκριση-evaluation με ένα manually selected σετ από όρους. Αξιοποιεί δέντρα αποφάσεων, predictors και διαδικασίες επιλογής κλαδιών στο δέντρο *(pruning)*.

Όλα τα πειράματα των άνωθι μεθόδων αξιολογήθηκαν με της κλασσικές μετρικές της ακρίβειας, ανάκλησης και F-measure. Τα συμπεράσματα που προέκυψαν ήταν ότι η συνδυαστική μέθοδος (που συμπεριλαμβάνει στατιστικές, γλωσσολογικές αλλά και ευριστικές μεθόδους) έχει οριακά καλύτερες επιδόσεις από τις προηγούμενες που υπήρχαν ως μέτρα σύγκρισης [15].

2.4 Νέες Προσεγγίσεις Αυτόματου Εντοπισμού Όρων

Στο paper Automatic term identification for bibliometric mapping [16] επιχειρείται μία βελτίωση στην αυτόματη αναγνώριση όρων ώστε να χρησιμοποιηθεί για την δημιουργία «βιβλιομετρικής χαρτογράφησης (bibliometric mapping)». Αυτή αποτελεί «έναν χάρτη που οπτικοποιεί τη δομή ενός επιστημονικού τομέα δείχνοντας τις σχέσεις ανάμεσα σε σημαντικούς του όρους» [16]. Πιο συγκεκριμένα, αφού αναλυθούν οι γνωστοί λόγοι που προτιμείται η αυτοματοποίηση από την χειροκίνητη εξαγωγή όρων, καθώς επίσης αναλυθούν οι ήδη υπάρχοντες co-word maps και co-occurrence maps (που αφορούν συσχετίσεις λέξεων και άρθρων αντίστοιχα), υποδηλώνεται ο στόχος δημιουργίας ενός term map που θα αφορά τους όρους ενός τομέα.

Πολύ σημαντικό είναι να τονιστεί ότι εκτός από την εύρεση όρων γενικά του τομέα, είναι συχνά θεμιτή μια κατηγοριοποίηση τους εντός των διαφόρων κλάδων (θέματα = topics) που συνιστούν τον επιστημονικό τομέα. Αυτό είναι ουσιώδες για να διαφανεί στον χάρτη η διακριτότητα των κλάδων, μιας και ένας όρος για έναν κλάδο δεν είναι όρος για έναν δεύτερο. Για παράδειγμα, ο όρος οπτικός φακός μπορεί να είναι όρος για έναν οφθαλμίατρο αλλά για τον κλάδο της κτηνιατρικής δεν αναμένουμε πολλαπλές αναφορές σε αυτόν. Η βάση για τον διαχωρισμό αυτόν είναι η υπόθεση ότι κάθε όρος αντιστοιχίζεται σε ένα μόνο θέμα. Επομένως, αν μια λέξη ή φράση εμφανίζεται πολλές φορές στα κείμενα ενός θέματος αλλά λίγες έως καθόλου σε αυτά όλων των άλλων θεμάτων, αυτό την καθιστά εξαιρετικά πιθανό όρο του πρώτου θέματος. Αντίστροφα, αν εμφανίζεται πολλές φορές σε όλα τα θέματα, είναι πιθανό να μην αποτελεί όρο κανενός θέματος, αλλά γενικότερη και μη εξειδικευμένη φράση. Πιθανή απόκλιση από την 1 προς 1 αντιστοιχία είναι εφικτή με μια τεχνική ονομαζόμενη Probabilistic Latent Semantic Analysis, η οποία συνοψίζεται στην υπόθεση ότι είναι πιθανότερη η εύρεση στο ίδιο κείμενο των λέξεων που είναι κοντά σημασιολογικά.

Επομένως, η τεχνική που εφαρμόζεται προς αυτόν τον σκοπό είναι πρώτα να εντοπιστούν τα διαφορετικά θέματα σε ένα σύνολο κειμένων, στη συνέχεια να ευρεθούν οι συχνές φράσεις και να σημαδευτούν ως προς τα μέρη λόγου που τις απαρτίζουν με βάση αποδεκτά μοτίβα, και στη συνέχεια να υπολογιστεί η (“unithood”) αλλά και η (“termhood”) των υποψήφιων όρων, με τη δεύτερη να εξαρτάται και από τον διαχωρισμό/μοναδικότητα της φράσης στα διάφορα θέματα όπως αναλύθηκε πριν. Μέσα από την Probabilistic Latent Semantic Analysis, εντοπίζονται τα ξεχωριστά θέματα, και εκτιμώνται με το κριτήριο του max log-likelihood.

Συνοψίζοντας, εκτιμάται ότι ο διαχωρισμός του termhood και του unithood μέσα από τον διαχωρισμό ενός corpus κειμένων σε διαφορετικά θέματα, εκτός του ότι εξυπηρετεί τους σκοπούς και τους λόγους

δημιουργίας ενός χάρτη όρων, αποδεικνύει ότι ενισχύει την ακρίβεια και την ανάκληση στην αυτόματη εξαγωγή όρων, με βάση και τα αποτελέσματα που αναδεικνύουν οι ειδικοί.

Σε ένα άλλο paper των Nenadic G, Ananiadou S και McNaught J [17] γίνεται αναφορά στη παραλλαγή των όρων (term variation) η οποία υποδηλώνει τις διαφορετικές λεκτικές μορφές που μπορεί να έχει η έκφραση της ίδιας έννοιας και ανάγει την εξαγωγή όρων σε ένα δυσκολότερο πρόβλημα ομαδοποίησης. Προτείνεται μια μέθοδος υβριδικής προσέγγισης τόσο με γλωσσολογικά όσο και με στατιστικά φίλτρα ώστε να λαμβάνεται υπόψιν η συνδυαστική συχνότητα όλων των μορφών και εμφανίσεων των παραλλαγών ενός όρου.

Δεδομένης της συχνότητας εμφάνισης παραλλαγών στους όρους, είναι αναγκαία η συσχέτιση και η ομαδοποίηση τους σε υπερσύνολα που εκφράζουν την ίδια έννοια με πολλές διακριτές μορφές. Αρχικά, διερευνώνται οι τρόποι που μπορεί ένας όρος να διαθέτει παραλλαγές, οι οποίοι είναι ορθογραφικοί, μορφολογικοί, λεκτικοί, δομικοί και μέσω ακρωνυμίων. (ορθογραφικά: amino acid and amino-acid, μορφολογικά: cellular gene and cell gene, λεκτικά: carcinoma and cancer, δομικά: clones of human and human clones, ακρωνύμια: DNA for deoxyribonucleic acid) Διαπιστώνεται ότι το stemming δεν επαρκεί σαν απλοϊκός τρόπος ομαδοποίησης διαφορετικών μορφών μιας λέξης, καθώς από μόνο του μπορεί να μην επαρκεί ή και να παράγει false positives ομαδοποιώντας όμοιες μορφολογικά λέξεις με τελείως διαφορετική ερμηνεία e.g. university, universe -> univers ενώ προέρχονται από τελείως διαφορετικούς όρους.

Γενικά, αιτιολογείται η ανάγκη ενός στάδιού κανονικοποίησης των ξεχωριστών υποψηφίων όρων και της αναγωγής τους σε έναν κανονικοποιημένο αντιπρόσωπο (“canonical representative”) [17]. Οι μέθοδοι που χρησιμοποιούνται είναι για παράδειγμα η επέκταση των POS μοτίβων ώστε να εξετάζουν την ύπαρξη προθέσεων ανάμεσα σε φράσεις που περιέχουν ίδιες λέξεις (π.χ. human cancer -> cancer in human), η χρήση λεξικών συνωνύμων με αναγωγή όλων των συνωνύμων σε έναν και μόνο αντιπρόσωπο, η αντιμετώπιση ορθογραφικών λαθών μέσα από επιλογή χαρακτήρων-γραμμάτων αντιπροσώπων αλλά και η παραγωγή και ομαδοποίηση πιθανών υποόρων ή και ακρωνυμίων μέσα από την δημιουργία όλων των πιθανών παραλλαγών που μπορούν να παραχθούν.

Τέλος, αφού εκτελεστούν όλες οι παραπάνω τεχνικές, ο υπολογισμός του termhood εκτελείται μόνο στους κανονικοποιημένους αντιπροσώπους, οι οποίοι θα είναι και οι τελικοί προτεινόμενοι όροι, μέθοδος που αυξάνει αποδεδειγμένα την ακρίβεια και την ανάκληση.

2.5 Αξιολόγηση αποτελεσμάτων Αυτόματου Εντοπισμού Όρων

Οι Korkontzelos Ioannis και Klapaftis I με το paper τους το 2008 [18] δίνουν έμφαση στην μελέτη διαφορετικών μεθόδων αυτόματης αναγνώρισης όρων και συγκεκριμένα επιλέγεται ο διαχωρισμός τους σε 2 μεγάλες κατηγορίες: Τις *termhood-based (εννοιολογικά βασισμένες)* προσεγγίσεις και τις *unithood-based (δομικά βασισμένες)* προσεγγίσεις. Αξίζει να σταθούμε λίγο παραπάνω στην οπτική τους για σύγκριση και αποτίμηση τους υπό ένα κοινό πρίσμα αξιολόγησης.

Η διαδικασία ξεκινά με την εφαρμογή ενός γλωσσολογικού φίλτρου, ενώ στη συνέχεια, εφαρμόζεται η εκάστοτε στατιστική μέθοδος. Τέλος, σύμφωνα με ένα «χρυσό πρότυπο»[2] (GENIA Gold Standard: ένα λεξικό όρων βιολογίας από το 2003, αναγνωρισμένο στον κλάδο του) αξιολογούνται τα κορυφαία αποτελέσματα της μεθόδου ώστε να υπάρξει συνολική αποτίμηση.

Στο άρθρο γίνεται ξεχωριστή αναφορά στους υπολογισμούς που εφαρμόζει κάθε στατιστική μέθοδος. Αποδεικνύεται ότι οι πρώτες μέθοδοι (termhood-based) είναι ακριβέστερες από τις δεύτερες (unithood-based), με σύγκριση διαφόρων εκπροσώπων όπως των *C-Value, Statistical Barrier (SB) (=σύνθετοι όροι αποτελούνται από υπάρχοντες απλούστερους όρους)* από τη μεριά της termhood, και των *Log-likelihood (LL), Pointwise Mutual Information (PMI)* από τη μεριά της unithood.

Αυτό φαίνεται να σχετίζεται με το γεγονός ότι οι δεύτερες μέθοδοι μετρούν την δύναμη της σύνδεσης των συνθετικών ενός υπο-όρου, χωρίς να σημαίνει ότι αυτοί καθαυτοί οι υποόροι ανήκουν στον τομέα μελέτης, όταν χρησιμοποιηθούν ανεξάρτητα. Δηλαδή, μία φράση μπορεί να αποτελείται από απλές λέξεις οι οποίες από μόνες τους δεν αποτελούν επιστημονική έννοια, αλλά όταν τεθούν με έναν εξαιρετικά συγκεκριμένο συνδυασμό κάποιες εξ' αυτών μαζί, συνιστούν έναν τεχνικό όρο. (π.χ. «contact lens», (ελ. οπτικός φακός), ο φακός δεν είναι όρος της ιατρικής επιστήμης ενώ contact, ως επαφή επίσης δεν αποτελεί όρο. Μαζί όμως αναφέρονται στους φακούς επαφής που αφορούν την οφθαλμιατρική. Η δομική αποτίμηση (unithood) θα θεωρήσει αυτές τις λέξεις πιθανούς όρους, ενώ η εννοιολογική αποτίμηση (termhood), σωστά, θα τις απορρίψει.

Έτσι, «το πλήθος των ανεξάρτητων εμφανίσεων ενός υποψήφιου όρου φαίνεται να είναι η πιο αποδοτική πηγή εμφωλευμένης πληροφορίας» [18]

Σε ένα άλλο paper [19] συγκρίνονται οι voting algorithms με τις μεθόδους μηχανικής μάθησης στον κλάδο της αυτόματης αναγνώρισης όρων. Αφού πρώτα εξηγηθεί η διαδικασία και τα βήματα που

[2] Όμοια και εμείς στο πειραματικό σκέλος της διπλωματικής αυτής εργασίας, θα αξιολογήσουμε τα αποτελέσματα μας με βάση την ήδη υπάρχουσα οντολογία HarmonicSS, που απασχολείται με το σύνδρομο Sjogren βλ. HarmonicSS Project https://www.harmonicss.eu/

ακολουθεί η αυτόματη αναγνώριση όρων, καθώς επίσης συνδεθούν οι νέες αυτές μέθοδοι με τις ήδη υπάρχουσες τεχνικές που εντάσσονται στις έννοιες του termhood και του unithood, προτείνεται συνδυασμός και των 2 κατηγοριών αλγορίθμων (μηχανικής μάθησης και ψήφου).

Μετά από το γλωσσολογικό και το στατιστικό βήμα, οι μέθοδοι ψήφου καταλήγουν να αξιολογούν κάθε πιθανό όρο με κάποιου είδους μετρική όπως η loglikelihood (unithood), domain relevance (termhood) και η C-Value(hybrid) καθώς και άλλες που παρουσιάζονται στο συγκεκριμένο paper. Προτείνεται από τους Fedorendko D, Astrakhantsev N, Turdakov D [19] να υπάρχει μία σταθμισμένη μέση τιμή η οποία θα εκπροσωπεί και θα συνοψίζει όλες τις άνωθι μετρικές ώστε να υπάρχει ενιαία αξιολόγηση, δεδομένου ότι έχουν αντίστοιχη αξία και «η τάση είναι να συνδυάζονται όλες μαζί». Όσον αφορά τη μηχανική μάθηση, αξιοποιούνται 2 μέθοδοι (*Random Forest and Logistic Regression)*, οι οποίες είναι ενδεικτικές «λόγω της αποτελεσματικότητας τους και της καλής γενίκευσης που παρέχουν στο παραγόμενο μοντέλο» και τα datasets χωρίζονται σε σύνολα εκπαίδευσης και σύνολα δοκιμής (10-90 αναλογία). Η μέθοδος είναι να βγάζουμε τις παραπάνω αναφερθείσες μετρικές ως features για το μοντέλο μηχανικής μάθησης, να τις μοιράζουμε ως datasets στις 2 άνω κατηγορίες και να επαναλαμβάνουμε τη διαδικασία εκπαιδεύοντας το μοντέλο μας. Τα inputs καταλήγουν να έρχονται από το GENIA Gold Standard ως training, ενώ το BIO1 αποτελεί το testing σύνολο.

Για την αξιολόγηση, οι τελικές μετρικές που αξιοποιούνται είναι η *ακρίβεια (precision)* που είναι ο λόγος των σωστών αποτελεσμάτων προς όλα τα επιστραμμένα αποτελέσματα και η *ανάκληση (recall)* που είναι ο λόγος όλων των σωστών αποτελεσμάτων προς όλα τα δυνατά σωστά αποτελέσματα που μπορεί να υπάρχουν. Με λίγα λόγια, ελέγχουμε το κατά πόσο το BIO1 ήταν κοντά στο GENIA, μέσω των μετρικών precision και recall.

Γενικά, συμπεραίνεται πειραματικά ότι οι μέθοδοι μηχανικής μάθησης είναι ισχυρότερες και ακριβέστερες από τις αντίστοιχες μεθόδους «ψήφου». Προτείνεται μία προσέγγιση αξιοποίησης μικρών συνόλων εκπαίδευσης για εκπαίδευση ενός *ταξινομητή (classifier)*, με τον οποίο θα εξάγονται νέοι όροι, μέρος (το βέλτιστο) εκ των οποίων θα επιστρέφονται για αναδρομική εκπαίδευση, με τη διαδικασία να επαναλαμβάνεται μέχρις ότου να υπάρχουν ασφαλή συμπεράσματα. Φυσικά, διαπιστώνεται ότι αυτή η μέθοδος υστερεί όταν το σετ δεδομένων είναι ετερογενές σε μεγάλο βαθμό, καθώς η κατανομή των όρων είναι διάσπαρτη και διαφορετική σε κάθε κείμενο.

2.6 Συσχέτιση με mapping και εφαρμογές σε μη αγγλικά κείμενα

Παραπλήσιες τεχνικές που θα μας οδηγούσαν στην αξιολόγηση των διαφορετικών μορφών με τις οποίες εμφανίζεται ένας τεχνικός όρος σχετίζονται με την επονομαζόμενη *«Οντολογική Ευθυγράμμιση» (Ontology Alignment)* [20]. Αυτή αποτελεί μία μέθοδο διασύνδεσης ανάμεσα σε διαφορετικές συμβολοακολουθίες και εκφράσεις μέσα από αντιστοιχίσεις μεταξύ των εννοιών και των σχέσεων τους με ημί-αυτοματοποιημένες τεχνικές, με στόχο να αντιστοιχιστούν οντολογίες. Ο πλήρης αυτοματισμός συχνά κρίνεται ανεπαρκής μιας και τα συνώνυμα και οι ομώνυμες λέξεις (π.χ. *bat: ρόπαλο αλλά και νυχτερίδα*) απαιτούν έλεγχο από πραγματικό χρήση. Έτσι, αξιοποιούνται οπτικοποιήσεις και περιβάλλοντα χρήστη, με βέλτιστο τρόπο, ώστε η συνεισφορά του ανθρώπου να δένει αρμονικά ως οδηγός και επαληθευτής των αυτόματων αποτελεσμάτων.

Η οντολογία είναι ένα αφηρημένο μοντέλο αναπαράστασης του πραγματικού κόσμου, το οποίο διέπεται από καθορισμένες έννοιες, κανόνες, σχέσεις και ιδιότητες και χρησιμοποιείται ως εξωτερικός πάροχος γνώσης με εξαιρετική αποδοτικότητα. Ως μοντέλα αναπαράστασης, μπορεί να αλληλεπικαλύπτονται αλλά και να διαφέρουν για την προσέγγιση του ιδίου τομέα επομένως απαιτείται μία κάποια *«Οντολογική Διαμεσολάβηση» (Ontology Mediation)* [20] η οποία συμπεριλαμβάνει μετά την ευθυγράμμιση τα στάδια της χαρτογράφησης και της συγχώνευσης σε μία ενιαία οντολογία.

Σχετικά με το στάδιο της ευθυγράμμισης μεταξύ των οντολογιών, διακρίνουμε δύο στάδια, πρώτα το αυτοματοποιημένο και στη συνέχεια το καθοδηγούμενο από τον άνθρωπο, τον ειδικό. Το δεύτερο στάδιο οφείλει να συνίσταται αρχικά ως μία οπτική αναπαράσταση η οποία προκύπτει από την αυτόματη διαδικασία αλλά και η οποία στη συνέχεια λειτουργεί ως ανατροφοδότηση στο στάδιο αυτοματοποίησης στην παραγωγή των σωστών σχέσεων μεταξύ των οντολογιών. Οι αντιστοιχίσεις ευθυγράμμισης μεταξύ των εννοιών ανάμεσα σε 2 διαφορετικές οντολογίες μπορούν να δοθούν ως μια τριπλέτα δεδομένων που περιλαμβάνει τα 2 κλειδιά των εννοιών στην οντολογία τους και 3^{ον} μία δυϊκή λογική μεταβλητή, η *«εμπιστοσύνη» (Confidence).*

Η άνωθι εμπιστοσύνη μπορεί να προκύπτει από μια ποικιλία μεθόδων που αποσκοπούν στην ευθυγράμμιση των οντολογιών. Για παράδειγμα, στο επίπεδο των συμβολοσειρών που συνιστούν τις έννοιες μπορούμε να τις συγκρίνουμε μορφολογικά με τη βοήθεια τεχνικών κανονικοποίησης (*tokenization*) δηλαδή αναγωγή των λέξεων σε μοναδικά κλειδιά που περιλαμβάνουν δειγματοληπτικά τμήμα (κυρίως τα σύμφωνα) μίας λέξεις. Επίσης, υπολογίζεται edit distance στις λέξεις, δηλαδή σε πόσα σύμβολα διαφέρουν οι υποψήφιες όμοιες λέξεις, με αποτέλεσμα να οδηγούμαστε έτσι σε μια κατά προσέγγιση αντιστοίχιση των συμβολοσειρών (*approximate string matching)*. Τέλος, αφού έχουν προηγηθεί μπορούν να ομαδοποιηθούν με μεθόδους αντίστοιχες με την τεχνική του clustering οι έννοιες

που βρίσκονται κοντά εννοιολογικά, ή και να εφαρμοστεί το ίδιο σε ακολουθίες λέξεων, τα λεγόμενα n-grams (μελετώντας δηλαδή πλέον φράσεις αντί για λέξεις). Υπάρχει επίσης η δυνατότητα αξιοποίησης εξωγενών γλωσσικών πόρων όπως βάσεις δεδομένων-λεξικά (Wordnet[3]) για εύρεση και αντιστοίχιση συνωνύμων, μέθοδοι βασισμένες σε περιορισμούς (Constraint-based methods) και μέθοδοι βασισμένες στην δομή της οντολογίας, οι οποίες ελέγχουν τις γειτονικές σχέσεις ανάμεσα σε υποψήφιες έννοιες και αν αυτές είναι πολύ όμοιες, συμπεραίνουν ότι και οι έννοιες είναι όμοιες. Όμοια υπάρχουν τεχνικές υβριδικές, σύνθετες και βασισμένες σε μηχανική μάθηση.

Ιδιαίτερη σημασία έχει η επιτυχής αξιολόγηση των άνω τεχνικών, η οποία λαμβάνει χώρα σε καλύτερες συνθήκες όταν υπάρχει άμεση συμμετοχή ανθρώπων στη διαδικασία. Υπάρχουν όμως αυστηρές προϋποθέσεις για να δομηθεί μια αξιόπιστη συνεργασία ανθρώπου-μηχανής και αυτές συνοψίζονται στις *απαιτήσεις καθορισμένες από τον χρήστη (User Driven Requirements),* και στις *απαιτήσεις καθορισμένες από τη διαδικασία (Process Driven Requirements)*. Οι πρώτες συνοψίζονται στην καλλιέργεια ενός κλίματος εύκολης συνεργασίας μεταξύ των χρηστών που αξιολογούν τα αποτελέσματα της μηχανής, αποτελεσματικής αναπαράστασης των αποτελεσμάτων αυτών από τη μηχανή στο χρήστη και γενικότερα μια ευθυγράμμιση μεταξύ του ανθρώπινου τρόπου εργασίας με την καθορισμένη διαδικασία της αυτοματοποίησης. Οι δεύτερες αφορούν γενικά τον τρόπο με τον οποίο η ανατροφοδότηση από τον χρήστη εντάσσεται εντός του αλγορίθμου αυτοματοποίησης και σε ποια σημεία η μηχανή προβλέπει την επέμβαση αυτή.

Στη βιβλιογραφία μας συμπεριλαμβάνουμε και μια απόπειρα των Conrado M, Pardo T και Rezende S [21] για αυτόματη εξαγωγή όρων στην πορτογαλική γλώσσα. Επεξηγούνται στην αρχή από τους συγγραφείς τα προβλήματα που προσπαθεί να αντιμετωπίσει γενικά ο συγκεκριμένος επιστημονικός τομέας και στη συνέχεια αναφέρονται στην συνεισφορά που μπορεί να παρέχει σε αυτό η μηχανική μάθηση. Αποτελεί μια καλή ευκαιρία για να διαπιστωθούν τα προβλήματα που θα κληθούμε να αντιμετωπίσουμε και σε μια δική μας απόπειρα ενός συστήματος αυτόματης αναγνώρισης λέξεων.

Με μία πρώτη ανάγνωση, διαπιστώνεται ότι τα σημαντικότερα θέματα στον τομέα αυτό αφορούν τις έννοιες του *«θορύβου»* και της *«σιγής»,* δηλαδή των false-positives και των non-positives, εννοιών δηλαδή που διαπιστώνονται ως όροι ενώ δεν πρέπει, ή όρων που δεν βρίσκονται. Επιπρόσθετα προβλήματα είναι ο υψηλός αριθμός υποψήφιων όρων, η μεγάλη διάρκεια ανθρώπινης επαλήθευσης των αποτελεσμάτων και η χαμηλή αποτελεσματικότητα των αυτόματων μηχανισμών εύρεσης όρων.

[3] Η online βιβλιοθήκη του wordnet: https://wordnet.princeton.edu/

Με την προσέγγιση που ακολουθούν οι ερευνητές της δημοσίευσης αυτής, επιχειρούν να αξιολογήσουν τα αποτελέσματα της συνδυαστικής εφαρμογής μεθόδων στατιστικής γλωσσολογικής και υβριδικής πληροφορίας για τους υποψήφιους όρους, με την αξιοποίηση πολλαπλών γνωστών μετρικών (Term Variance, Term Cariance Quality, Term Contribution, Machine Learning). Στην αρχική φάση της μεθόδου τους, αξιοποιούν *κανονικοποίηση* των λέξεων των κειμένων με ταυτόχρονη *Επισήμανση Μέρους του Λογου (POS tagging)*. Η μελέτη τους γίνεται στην πορτογαλική γλώσσα και βασικό στόχο έχει να μελετήσει την βελτίωση των τεχνικών Αυτόματης Εξαγωγής όρων δια μέσου της ανάμιξης στατιστικών γλωσσολογικών και υβριδικών μεθόδων τόσο με δεδομένη είσοδο, όσο και με εξωτερικά σώματα κειμένων. Πράγματι, μέσα από τη δημιουργία 4 διαφορετικών δοκιμών με διαφορετική ανάμιξη των μετρικών, βελτιώθηκε η ακρίβεια, το F-measure και ελαχιστοποιήθηκε ο θόρυβος και η σιγή.

3 Μεθοδολογία

3.1 Ο στόχος της εφαρμογής

Στο πλαίσιο αυτό της τρέχουσας επιστημονικής τεχνογνωσίας που παρουσιάσαμε πρωτύτερα, προτείνουμε τη δημιουργία ενός προγράμματος το οποίο θα λειτουργεί με τις αρχές αλγορίθμων υβριδικής (στατιστικής αλλά και γλωσσολογικής) ανίχνευσης όρων. Στόχος μας είναι με την εφαρμογή αυτή να αξιοποιήσουμε τον αλγόριθμο του C-Value ως πηγή εύρεσης ιατρικών εξειδικευμένων όρων που σχετίζονται με το σύνδρομο Σιόγκρεν ώστε να μπορεί να χρησιμοποιηθεί για τον εμπλουτισμό ήδη υπαρχουσών οντολογιών, με συστάσεις προς τους ιατρικούς ερευνητές για ενδεχόμενους όρους. Επιπλέον, με την εφαρμογή αυτή στοχεύουμε να μπορεί να χρησιμοποιηθεί στην εξαγωγή της κρυμμένης πληροφορίας μέσα στις νοσοκομειακές σημειώσεις, με εστίαση στους σημαντικότερους όρους που χρησιμοποιούνται.

Το αντικείμενο μελέτης και εφαρμογής του παραπάνω προγράμματος που θα αναπτύξουμε θα αφορά δεδομένα σχετικά με το σύνδρομο Σιόγκρεν, με στόχο τον εμπλουτισμό ενός ήδη υπάρχοντος Μοντέλου Αναφοράς με όρους που θα ανιχνεύσουμε από τρεις (3) πηγές σχετικές με αυτό.

1) Η πρώτη εμπεριέχει σε μορφή xls μοντέλα περιγραφής της δομής των δεδομένων ασθενών με σύνδρομο Σιόγκρεν σε διαφορετικά Ινστιτούτα.

2) Η δεύτερη αφορά κλινικές δοκιμές που έχουν ήδη γίνει για το σύνδρομο Σιόγκρεν, οι οποίες έχουν δημοσιευτεί στο clinicaltrials.gov και εξήχθησαν με τη μορφή XML εγγράφων.

3) Η τρίτη αφορά επιστημονικά άρθρα και δημοσιεύσεις από online βιβλιογραφία για τον COVID-19 (CORD19 dataset), διαθέσιμα σε μορφή CSV και JSON και τα οποία έχουν φιλτραριστεί ώστε να ελεγχθεί ότι αναφέρεται εντός τους το σύνδρομο Σιόγκρεν.

Εκτός από τις άνωθι εφαρμογές, που θα εκτελεστούν υπό την μορφή πειραμάτων, χρειάστηκε πρώτα να αξιολογήσουμε κατά πόσο ο αλγόριθμος είναι αποδοτικός μέσα από την εκτέλεση μίας δοκιμής με χρήση ενός απλού και ελεγμένου dataset (στην περίπτωση μας η σελίδα του Sjogren Syndrome στην Wikipedia[4]), με ένα εξίσου απλού σετ γνωστών όρων (επιλέχθηκαν manually από εμάς) ούτως ώστε

[4] https://en.wikipedia.org/wiki/Sj%C3%B6gren_syndrome

να εντοπιστούν bugs και να εκτελεστούν βελτιώσεις μέχρι να έρθουμε σε επιθυμητό βαθμό απόδοσης του συστήματος.

3.2 Λειτουργία-Αρχιτεκτονική

Η προσέγγιση που θα ακολουθήσουμε στηρίζεται σε μία ιεραρχική ανάλυση της διαδικασίας καταγραφής και εντοπισμού των όρων σε επιμέρους υποπροβλήματα, για τα οποία θα υπάρξει ξεχωριστή υλοποίηση κώδικα και προγράμματος. Εντοπίζουμε λοιπόν τα εξής:

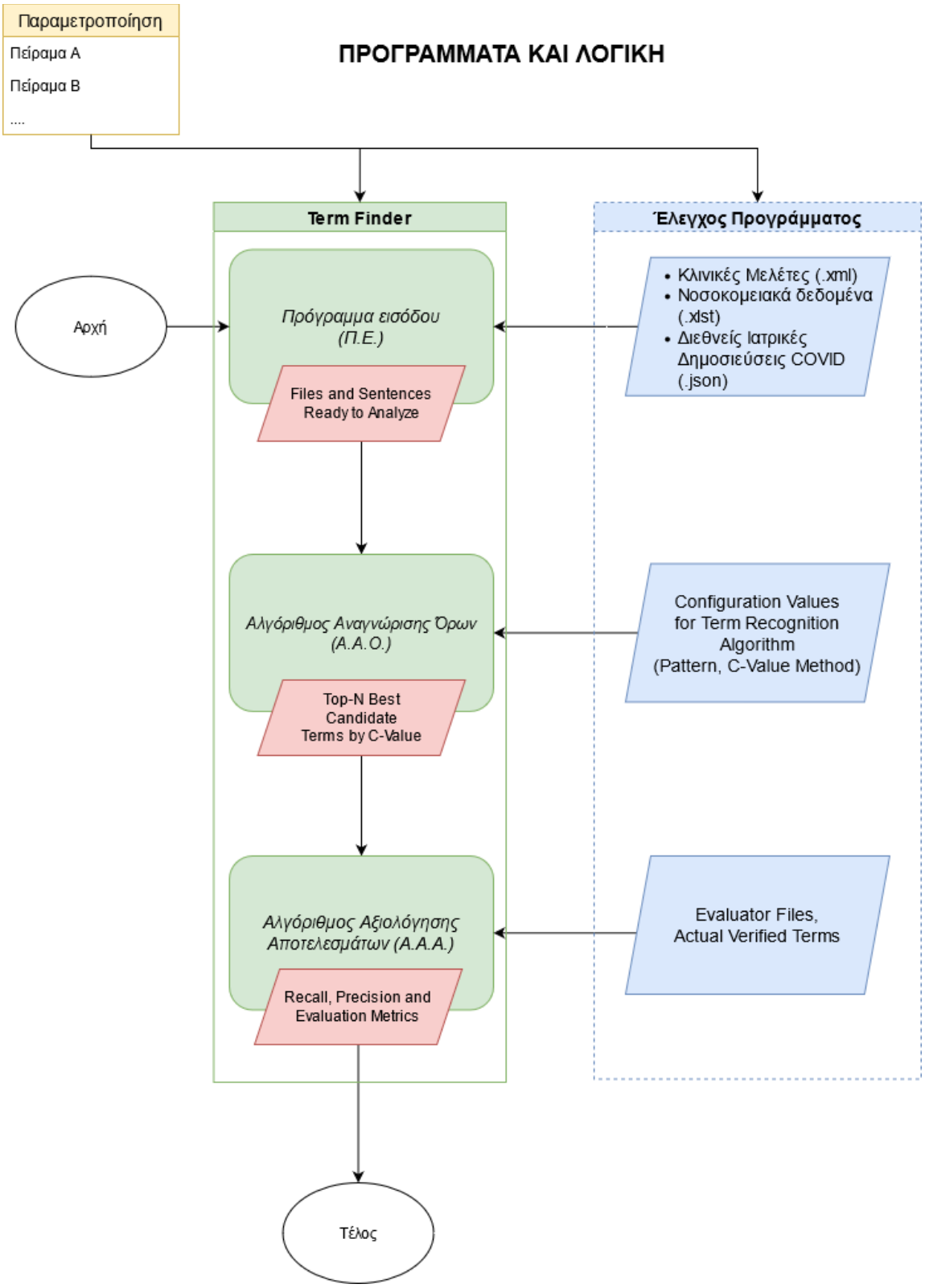
1) Ένα *πρόγραμμα εισόδου* για κάθε διαφορετική πηγή *(Π.Ε.)*. Στόχος του είναι να τα φέρει σε μια κανονικοποιημένη κοινή μορφή (Directory of .txt files) με συγκεκριμένη εσωτερική δομή (One sentence per line) ώστε να είναι κατάλληλη η μαζική εισαγωγή τους στον κεντρικό αλγόριθμο.

2) Ο *κεντρικός αλγόριθμος αναγνώρισης όρων (Α.Α.Ο.).* Στόχος του είναι με είσοδο τις προτάσεις, να εκτελέσει διαδικασίες stemming, tokenization διάφορες στατιστικές μετρικές και αλγορίθμους που ποσοτικοποιούν το κατά πόσο λέξεις ή φράσεις, στις διάφορες μορφές στις οποίες ομαδοποιούνται, αποτελούν όρους (termhood, unithood, C-Value).

3) Ο *αλγόριθμος αξιολόγησης των αποτελεσμάτων (Α.Α.Α)* . Στόχος του είναι να αποφασίσει για κάθε προτεινόμενο όρο που προέρχεται ως αποτέλεσμα του αλγορίθμου αναγνώρισης όρων (Α.Α.Ο.) κατά πόσο αποτελεί ή όχι πραγματικό τεχνικό όρο. Χρησιμοποιείται τόσο για τη δοκιμή και αξιολόγηση του συστήματος, όσο και για την πειραματική εφαρμογή του. Για τις ανάγκες της αξιολόγησης, οι αναμενόμενοι όροι καθορίστηκαν «με το μάτι» από το μικρό δοκιμαστικό κείμενο εισόδου, και δόθηκαν σε ένα txt file ως evaluator. Μελλοντικά, θα έπρεπε να γίνει με εισαγωγή του κειμένου σε γνωστές και διαπιστευμένες βάσεις δεδομένων (GENIA), ελεγμένες από ειδικούς του τομέα (domain experts).Με βάση την οντολογία του HarmonicSS [4] project για ιατρικούς όρους, εντοπίζουμε γνωστούς όρους στα κείμενα, αλλά και λαμβάνουμε άγνωστους όρους που βγαίνουν με ψηλότερο termhood από τους γνωστούς. Με αυτούς, φιλοδοξούμε να δώσουμε προτάσεις για διεύρυνση της οντολογίας μας.

Πίσω από την εκτέλεση του αλγορίθμου, προτού ξεκινήσει το κύριο μέρος, έχουμε το πρόγραμμα διαχείρισης υποσυστημάτων. Είναι σημαντικό να αναφερθούμε σε αυτό καθώς περιλαμβάνει και τη διεπαφή του χρήστη και αυτοματοποιεί τις πολλαπλές εκτελέσεις του αλγορίθμου που χρειαστήκαμε για τα πειράματα μας.

Οι σχέσεις μεταξύ των επιμέρους προγραμμάτων, οι είσοδοι και έξοδοι καθώς και οι παράπλευρες συνδέσεις με εξωτερικές πηγές περιγράφονται στο παρακάτω flowchart:



Σχήμα 2: Flow Chart για την αρχιτεκτονική του TermFinder

Είναι σημαντικό να αναφερθούμε αναλυτικότερα στα επιμέρους τμήματα του Termfinder:

Σχετικά με το πρόγραμμα εισόδου (Π.Ε.), αυτό λειτουργεί πρώτα ως ένας text preprocessor και έπειτα ως ένας text parser. Αρχικά στο στάδιο του preprocessor, αναγνωρίζει το περιεχόμενο του κειμένου (json, csv, txt, xls). Αφού αναγνωρίσει το πραγματικό σώμα κειμένου (body text) αποπειράται να εκτελέσει λειτουργίες tokenizer για να διαχωρίσει τις λέξεις και τις προτάσεις. Βρίσκοντας το σύμβολο της «.» αποπειράται να χωρίσει τα κείμενα σε γραμμές προτάσεων. Τέλος, εξάγει για κάθε αρχείο input αυτό το σετ προτάσεων ως ένα .txt αρχείο με κάθε γραμμή του να έχει μία πρόταση και το οδηγεί σε ένα directory.

Σχετικά με το πρόγραμμα του αλγορίθμου αναγνώρισης όρων (Α.Α.Ο.), η λειτουργία του είναι η υλοποίηση του αλγορίθμου C-Value που παρουσιάσαμε στο κεφάλαιο 2, με όλα τα επιμέρους βήματα. Εκτελεί POS-Tagging των tokens, με στόχο την εύρεση των φράσεων που ακολουθούν ένα συγκεκριμένο *μοτίβο* ακολουθίας μερών λόγου το οποίο δίνεται ως είσοδος, εύρεση διαφορετικών μορφών του ίδιου όρου, υπολογισμός C-Value υποψήφιων όρων με μία ή παραπάνω λέξεις κ.α. Το output που αφορά το termhood value για κάθε πιθανό όρο δίνεται στο τέλος με τη μορφή ενός text file στο οποίο φαίνονται με φθίνουσα σειρά ως προς τη μετρική αυτή οι υποψήφιοι όροι.

Σχετικά με το πρόγραμμα του αλγορίθμου αξιολόγησης αποτελεσμάτων (Α.Α.Α.), η λειτουργία του είναι να εξακριβώσει το κατά πόσο τα αποτελέσματα του αλγορίθμου αναγνώρισης όρων είναι έγκυρα και ακριβή. Ο στόχος του είναι διπλός: πρώτον, θα αξιολογήσει ως προς το ποσοστό των «γνωστών» όρων που πράγματι εντοπίζει σε δοκιμαστικά κείμενα για τα οποία γνωρίζουμε εκ των προτέρων τους όρους που θέλουμε να βρούμε. Δεύτερον, θα βοηθήσει στον εντοπισμό νέων όρων εν συγκρίσει με ήδη γνωστών όρων από κάποια οντολογία εισόδου. Η επιλογή των γνωστών όρων για την περίπτωση που πειραματιζόμαστε για την εύρεση αγνώστων όρων σε επιστημονικά κείμενα γίνεται από την οντολογία HarmonicSS [4].

Σχετικά με το πρόγραμμα διαχείρισης υποσυστημάτων, η λειτουργία του είναι να φορτώσει στα προαναφερθέντα προγράμματα το επιθυμητό Αρχείο Διαμόρφωσης, που περιέχει τις παραμέτρους που μας δίνουν έλεγχο στον τρόπο λειτουργίας του αλγορίθμου. Για παράδειγμα, μπορεί να εμπεριέχει το μοτίβο του POS-Tagging, τα αρχεία Evaluator, την εκδοχή του C-Value Algorithm και άλλα.

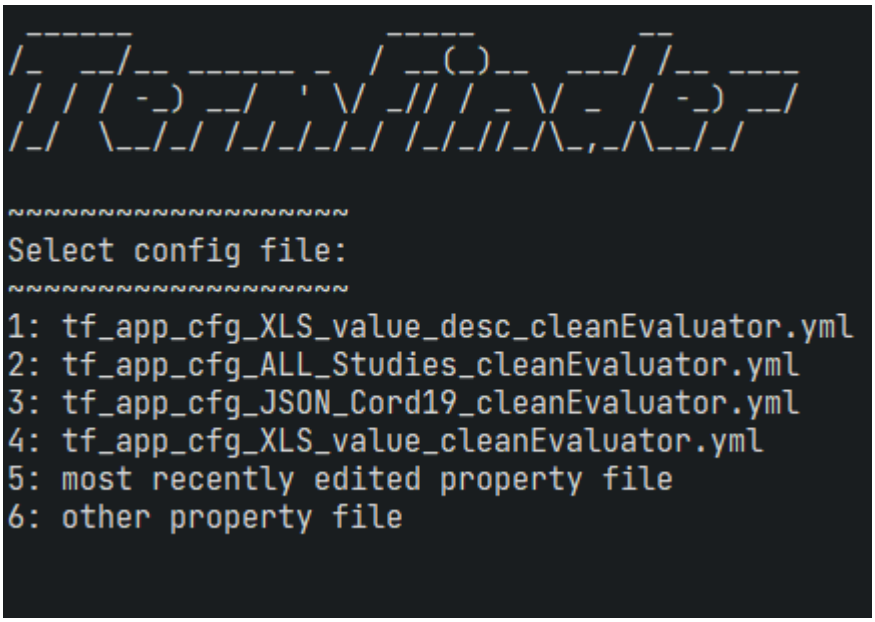
3.3 Τεχνικές λεπτομέρειες αλγορίθμου TermFinder

3.3.1 Πρόγραμμα Διαχείρισης Υποσυστημάτων (Integrator)

Το πρόγραμμα διαχείρισης υποσυστημάτων αποτελεί την πρώτη κλάση την οποία χρησιμοποιούμε για την εκτέλεση του αλγορίθμου. Περιλαμβάνει τη σειριακή διασύνδεση όλων των προγραμμάτων που θα αναφέρουμε στη συνέχεια και μπορούμε να πούμε ότι είναι ο «σκελετός» της εφαρμογής μας.

Περιλαμβάνει επιπλέον την console διεπαφή με τον χρήστη, η οποία στην τωρινή έκδοση της εφαρμογής μας αποτελεί ένα απλό οδηγό με Μηνύματα Οθόνης (terminal) που καθοδηγεί τον χρήστη να εισάγει το path για το configuration file με το οποίο θα τρέξει ο αλγόριθμος. Αυτό πρόκειται για το επονομαζόμενο Αρχείο Διαμόρφωσης και αποτελεί το ευρετήριο τόσο για εμάς όσο και για το λογισμικό όλων των παραμέτρων που υποστηρίζουν αλλαγές.

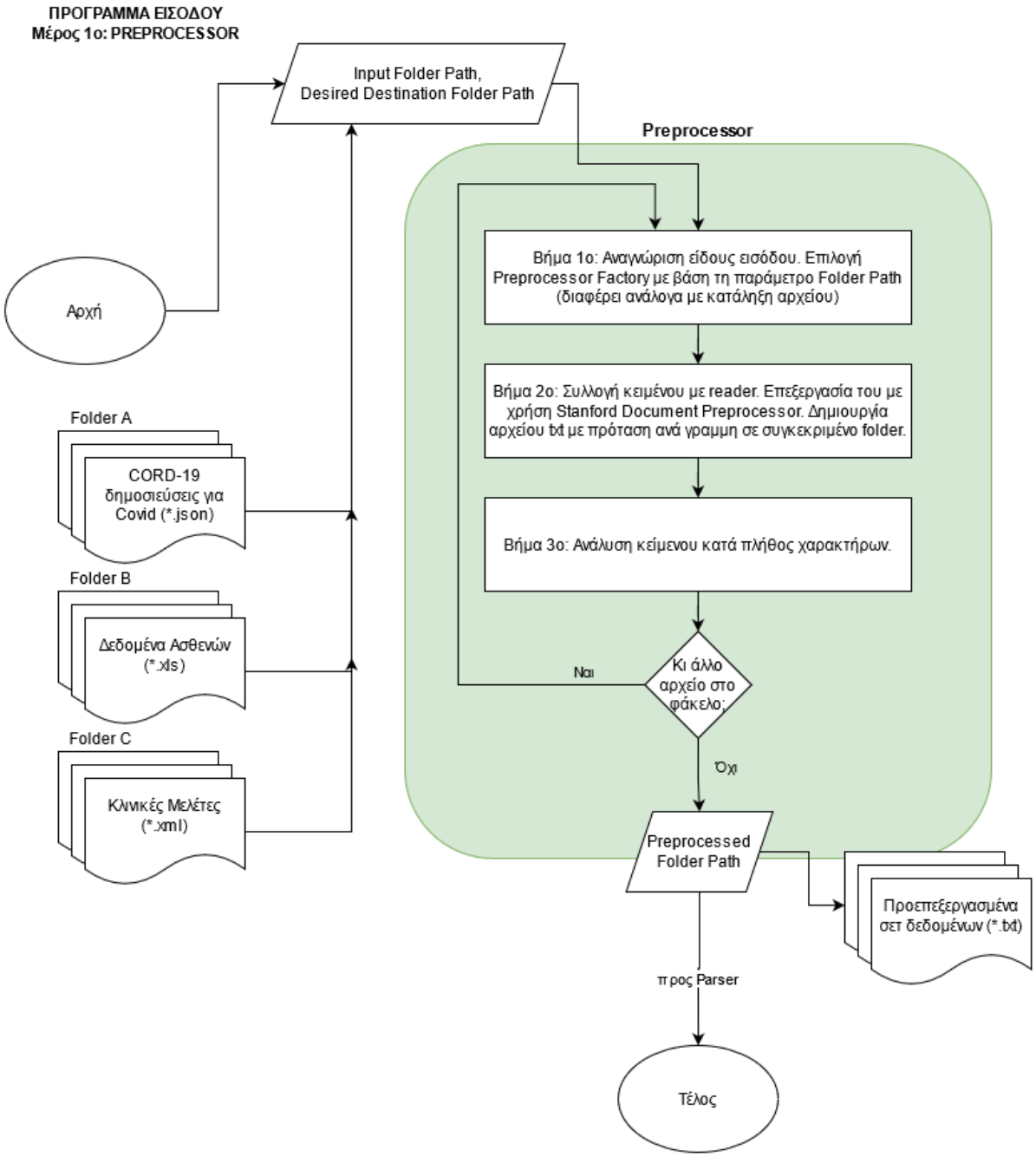
Για να μην υπάρχει περιττή εκτέλεση κώδικα, ο αλγόριθμος ελέγχει αν κάποια από τα configuration που αφορούν το evaluation, όπως για παράδειγμα το path για το *αρχείο αναζητούμενων πραγματικών όρων,* δεν χρειάζονται επανεκτέλεση όλου του αλγορίθμου, αλλά μπορούν να χρησιμοποιήσουν το output προηγούμενων πειραμάτων, επαναλαμβάνοντας μόνο τα επόμενα βήματα από αυτά που επηρεάζει η νέα διαμόρφωση.



Σχήμα 3: Διεπαφή χρήστη του TermFinder, πρόγραμμα διαχείρισης υποσυστημάτων.

3.3.2 Πρόγραμμα Εισόδου

Το πρόγραμμα εισόδου αποτελεί την πραγματική αφετηρία του συνολικού αλγορίθμου καθώς διαχειρίζεται την τροφοδοσία του *Termfinder* με τα αρχεία εισόδου με τα οποία θα τροφοδοτήσουμε τον αλγόριθμό μας. Απαρτίζεται από 2 μέρη, τον Preprocessor και τον Parser, με το πρώτο να τροφοδοτεί το δεύτερο σειριακά. Ξεκινώντας την εφαρμογή, ενεργοποιείται πρώτα ο preprocessor. Η λειτουργία του είναι ως εξής:



Σχήμα 4: Μέρος 1ο Προγράμματος Εισόδου (Π.Ε.) - Preprocessor

3.3.2.1 Πρόγραμμα Εισόδου - Preprocessor

Σε πρώτη φάση, το πρόγραμμα παραλαμβάνει τις αρχικές εισόδους μέσα από τη διεπαφή προεπεξεργασίας (preprocess interface). Για κάθε διαφορετικό είδος κειμένου, έχει αναπτυχθεί ένας εξαγωγέας και αναλυτής κειμένου (text preprocessor), με κοινή έξοδο για κάθε αρχείο ένα ξεχωριστό αρχείο το οποίο περιέχει σε κάθε γραμμή από μία πρόταση του κειμένου. Για να λειτουργήσει, τροφοδοτείται από το *αρχείο διαμόρφωσης* μία συμβολοσειρά με το μονοπάτι του folder (path) της εισόδου. Με βάση τις παραμέτρους που έχουμε επιλέξει, όπως για παράδειγμα την στήλη στο excel file που θέλουμε να διατρέξουμε, το πρόγραμμα εισόδου επεξεργάζεται όλα τα αρχεία μέσα στο path αναλόγως.

Οι περιπτώσεις εισόδων που συναντήσαμε είναι οι εξής:

1) .JSON Files: Σε αυτήν την περίπτωση χρησιμοποιούμε την org.json βιβλιοθήκη της java προκειμένου να αναζητήσουμε αναδρομικά κάθε value προερχόμενο από key με συγκεκριμένο όνομα. Εισάγουμε το κείμενο γραμμή-γραμμή στον Αναλυτή Κειμένου (Stanford DP) και σχηματίζουμε τα preprocessed txt, με αντιστοιχία 1 txt file για κάθε 1 json file.

2) .XLS Files: Σε αυτήν την περίπτωση με χρήση της βιβλιοθήκης org.apache.poi διακρίνουμε τα workbooks και τις στήλες από τις οποίες επιθυμούμε να μαζέψουμε κείμενο από το excel workbook και αναδρομικά για κάθε γραμμή λαμβάνουμε το περιεχόμενο, και όμοια με το json, μετά από τον Αναλυτή Κειμένου δημιουργεί 1 txt file για κάθε 1 excel file.

3) .TXT Files: Σε αυτήν την περίπτωση, απλά βάζουμε στον Stanford DP κάθε γραμμή του κειμένου μας και μετά για κάθε αρχείο χτίζουμε 1 νέο txt file με 1 πρόταση ανά γραμμή.

Είναι σημαντικό να διαχωρίσουμε τα Αρχεία Εισόδου από το *αρχείο διαμόρφωσης* που αναφέρθηκε στην προηγούμενη ενότητα, μιας και αυτά αποτελούν τα πραγματικά ανεπεξέργαστα δεδομένα τα οποία μπορούν να μας παρέχουν οι γιατροί, οι ερευνητές ή άλλοι αρμόδιοι του ιατρικού τομέα ώστε εμείς οι μηχανικοί λογισμικού να τα δώσουμε στον *Termfinder* να τα αντιληφθεί (μέσα από τις κατάλληλες οδηγίες στο *αρχείο διαμόρφωσης*, όπως paths, parameters και άλλα).

Στη συνέχεια, ακολουθεί η προσπέλαση των χαρακτήρων του κειμένου με βάση τα αναγκαία εργαλεία ανά περίπτωση ώστε να βρεθεί η χρήσιμη εξεταζόμενη πληροφορία. Για παράδειγμα, στην περίπτωση των Json Files, τα οποία έχουν δεντρική μορφή, γίνεται αναδρομική διάσχιση του δένδρου προς αναζήτηση values που αντιστοιχίζονται στο κλειδί με λεκτικό «body_text», καθώς εκτός από το κυρίως

κείμενο υπάρχουν επιπρόσθετες πληροφορίες για τον τίτλο και τις πηγές π.χ. τις οποίες θέλουμε να εξαιρέσουμε από την ανάλυση.

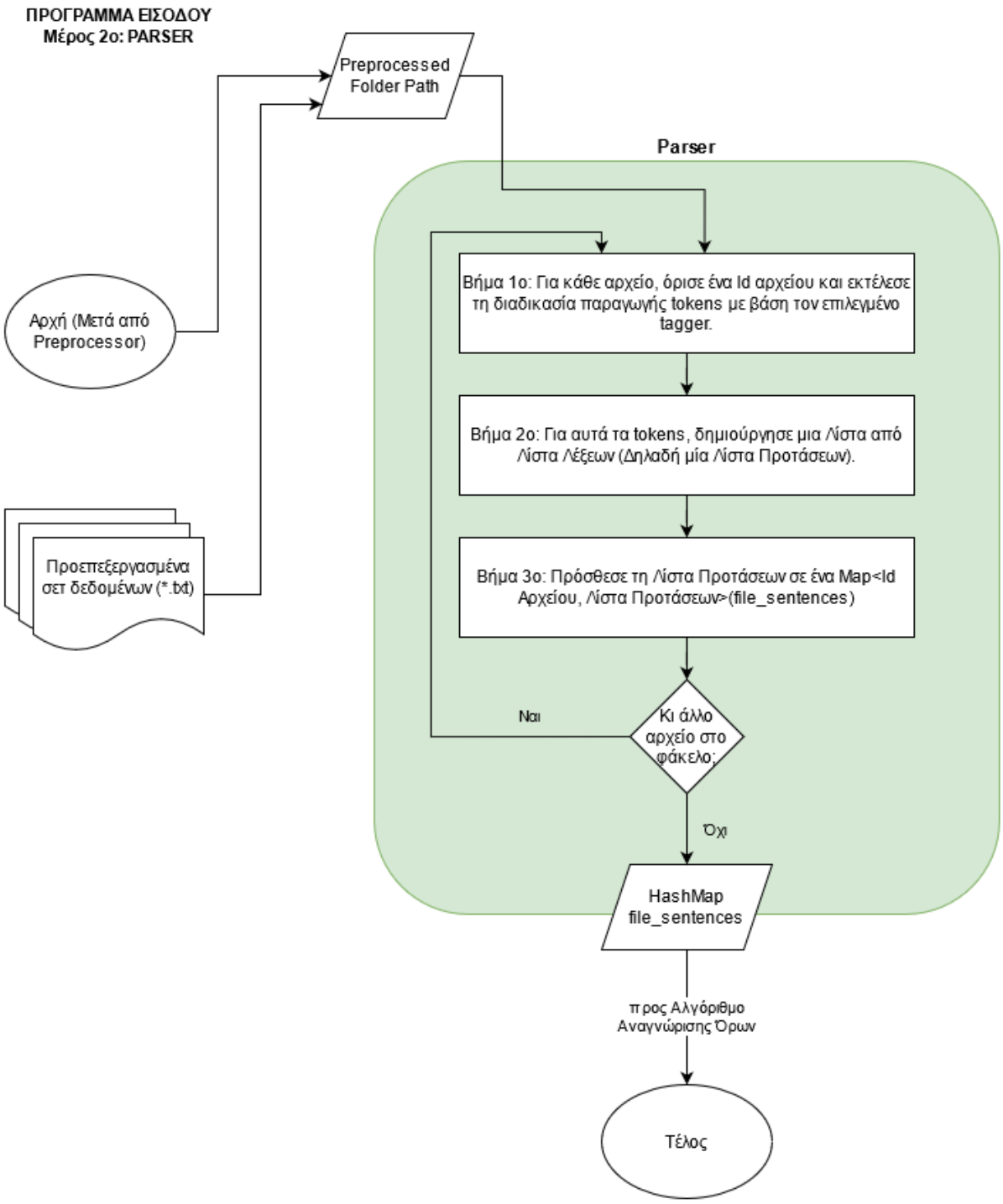
Μετά, έρχεται η σειρά για τη χρήση του Stanford Document Preprocessor. Η χρήση του αποσκοπεί στο να βρεθούν οι προτάσεις εντός του κειμένου και να διαχωριστούν κατά σειρά. Δίνει τη δυνατότητα για διαχωρισμό tag σε POS-Tagging, επιλεγεί άλλου delimiter, splitting και άλλες λειτουργίες αλλά σε αυτό το στάδιο μας αρκεί μόνο ο διαχωρισμός των προτάσεων.

Μόλις ολοκληρωθούν αυτά, για κάθε αρχείο συντάσσεται ένα ισοδύναμο txt αρχείο το οποίο περιέχει σε κάθε γραμμή του μία και μόνο μία πρόταση από το original text, ενώ επίσης περιέχει εξαντλητικά όλη τη χρήσιμη πληροφορία, είτε πρόκειται π.χ. για το body text των json files, είτε για συγκεκριμένα columns σε xls files. Αυτό το αρχείο αποθηκεύεται στον επιλεγμένο από την είσοδο φάκελο εξόδου και η διαδικασία επαναλαμβάνεται μέχρι να τελειώσουν τα αρχεία από τον φάκελο εισόδου.

Τέλος, γίνεται μία προεπεξεργασία και ανάλυση του συνόλου του dataset σε επίπεδο χαρακτήρων. Εντοπίζεται ποιοι χαρακτήρες υπάρχουν και πόσες φορές εμφανίζονται εντός όλων των κειμένων, ώστε να διαπιστωθεί αν χρειάζεται να γίνει κάποιο πρωταρχικό φιλτράρισμα στο κείμενο (π.χ. χαρακτήρες από άλλη γλώσσα εκτός από ελληνικούς και λατινικούς – αν τα κείμενα μας περιείχαν μεγάλο αριθμό λόγου χάρη αραβικών χαρακτήρων θα έπρεπε να τους εξαιρέσουμε ή να ελέγξουμε την είσοδο μας). Στις περιπτώσεις εισόδων μας κάτι τέτοιο δεν χρειάστηκε, καθώς το stoplist γίνεται σε επίπεδο λέξεων σε πιο ύστερο στάδιο. Μετά από τη λειτουργία του preprocessor, τα κανονικοποιημένα σετ δεδομένων είναι έτοιμα για parsing από τον tagger μας ώστε να ανιχνευθούν οι λέξεις και οι προτάσεις μέσα σε αυτά.

3.3.2.2 Πρόγραμμα Εισόδου - Parser

Έτσι, με είσοδο τον φάκελο όπου αποθηκεύτηκε το output του πρώτου μέρους του προγράμματος εισόδου, η εκτέλεση της εφαρμογής μας προχωρά στο 2ο στάδιο, αυτό του parser, με στόχο να προετοιμάσει την είσοδο για την εκτέλεση του αλγορίθμου. Συνοπτικά, η λειτουργία του μπορεί να αναπαρασταθεί με το παρακάτω διάγραμμα:



Σχήμα 5: Μέρος 2ο Προγράμματος Εισόδου (Π.Ε.) - Parser

Το πρώτο στάδιο αφορά τη σύνδεση με τον preprocessor, ή οποία γίνεται με την λήψη του path στο οποίο αποθήκευσε το output του. Εκεί, δεσμεύει σειριακά κάθε αρχείο το οποίο είναι ένα text file που αντιστοιχίζεται 1 προς 1 με κάποιο αρχείο που εμπεριεχόταν στο input dataset μελέτης. Κάθε γραμμή του, με τη σειρά του, αντιστοιχίζεται σε μία πρόταση και κάθε πρόταση αποτελείται από λέξεις. Ο

διαχωρισμός αυτός σε προτάσεις και λέξεις είναι χρήσιμος για το στάδιο του POS Pattern Matching που θα γίνει αργότερα, καθώς περιορίζει το εύρος αναζήτησης των μοτίβων που απαρτίζουν τους όρους σε συγκεκριμένο μέγεθος δείγματος.

Μετά από την είσοδο, ο αλγόριθμος του parsing εκτελεί για αυτό το αρχείο τα παρακάτω βήματα με στόχο να δημιουργήσει ένα HashMap με όλη την πληροφορία που θα τροφοδοτήσει στον Αλγόριθμο Αναγνώρισης Όρων: Πρώτα, ταυτοποιεί το αρχείο με ένα μοναδικό id και αρχικοποιεί τον δοσμένο tagger. Για κάθε γραμμή-πρόταση του αρχείου, με τη βοήθεια ενός lexical analyzer ξεκινά τη διαδικασία του tokenization με βάση τις αρχές του Penn Treebank Tokenizer. Πρακτικά, πέρα του ότι επιτυγχάνει να ανιχνεύσει τα διακριτά token-words, διαχωρίζει τα σημεία στίξης από τα άλλα tokens και μετατρέπει κάποιες συχνές συμπτύξεις της αγγλικής γλώσσας στις πρωταρχικές λέξεις.

Στη συνέχεια, τα tokens που παράγονται για κάθε πρόταση πρέπει να οργανωθούν και να ερμηνευθούν ως «λέξεις». Αυτές οι διακριτές λέξεις κάθε προτάσεως ερμηνεύονται από τον tokenizer με το ορισμένο object HasWord, τα οποία τοποθετούνται σε μία λίστα. Αυτή τη λίστα την οποία θα ονομάζουμε πλέον «πρόταση» δεν είναι άλλο από μία λίστα από HasWord. Αυτή η πρόταση με τη σειρά της οργανώνεται σε μία δεύτερη λίστα που είναι η «λίστα προτάσεων» μας, για αυτό το αρχείο.

Σειρά έχει να επαναλάβουμε τη διαδικασία αυτή για κάθε γραμμή του τρέχοντος αρχείου με στόχο να συμπληρώσουμε όλη τη λίστα προτάσεων εξαντλητικά για το αρχείο εισόδου μας. Το iteration ολοκληρώνεται για το αρχείο αυτό και ξεκινά το τελευταίο iteration που αποτελεί μία λίστα «λιστών προτάσεων», ή αλλιώς μία λίστα αρχείων η οποία εμπλουτίζεται από όλες τις προτάσεις κάθε αρχείου μαζί με ένα διακριτό αριθμό ως κλειδί του αρχείου. Είναι δηλαδή μία tuple της μορφής: Map<Id, List<List<HasWord>>>.

Αυτό θα το ονομάσουμε file_sentences και πάνω σε αυτή τη δομή δεδομένων θα γίνει όλη η επεξεργασία και η υλοποίηση του αλγορίθμου c-value. Περιέχει εξ'ολοκλήρου όλη την πληροφορία εισόδου δομημένα μεν, αλλά επί της οποίας δεν έχει γίνει ακόμη η περισυλλογή όρων. Όμως, είναι πλέον σε μορφή κατάλληλη για την εφαρμογή pattern matching αλγορίθμων, αφού έχει προηγηθεί το tokenization, ενώ το mapping μας δίνει την ευκαιρία να παρακολουθήσουμε και να διατρέξουμε με ασφάλεια όλη την έκταση του κειμένου.

Έτσι, ολοκληρώνεται η λειτουργία του προγράμματος εισόδου και ακολουθεί στη συνέχεια το κυρίως πρόγραμμα μας.

3.3.3 Αλγόριθμος Αναγνώρισης Όρων

Σε αυτή την ενότητα θα σταθούμε πιο διεξοδικά στα επιμέρους βήματα των αλγορίθμων και ιδιαίτερα στον κεντρικό αλγόριθμο αναγνώρισης όρων *(A.A.O)* ο οποίος είναι η «καρδιά» του *TermFinder*. Στόχος μας με τον αλγόριθμο αυτό, όπως προαναφέρθηκε, είναι από το κείμενο που του τροφοδοτείται να εντοπίσουμε λεκτικούς σχηματισμούς που είναι πιθανόν να αποτελούν όρους και να αναπαραστήσουμε αυτή την πιθανότητα με μετρήσιμη έκφραση μέσα από τη μετρική του C-Value.

Η διαδικασία εξέλιξης του αλγορίθμου πέρασε από 2 φάσεις.

Στην πρώτη επιχειρήθηκε η πιστή παρουσίαση της προσέγγισης του αλγορίθμου C-Value ως μεθόδου εύρεσης πιθανών όρων και αξιολόγησης τους μέσα από τον υπολογισμό του termhood τους.

Στη δεύτερη φάση, επιχειρήθηκε να βελτιωθεί η προσέγγιση αυτή μέσα από την ένταξη επιμέρους επιπρόσθετων κανόνων ώστε να αυξηθεί η ακρίβεια και να επιλυθούν τα διάφορα προβλήματα που παρουσιάσαμε νωρίτερα. Πέρα από την καθαρή εφαρμογή του αλγορίθμου, στο πρόγραμμα μας εντάξαμε βελτιώσεις στο γλωσσολογικό επίπεδο (stemming, term normalization, term variation, grammatical transformation), παραλλαγές του στο στατιστικό επίπεδο (“term bag” approach on multiword terms), υποβοήθηση από σημασιολογικής σκοπιάς στην ομαδοποίηση των όρων (εύρεση συνωνύμων μέσα από dictionaries) και τελικά υπολογίσαμε την επίδραση των άνωθι επεμβάσεων μέσα από πειράματα στα οποία μεταβάλλαμε τις παραμέτρους.

Η αξία του termhood ή αλλιώς C-Value αναλύθηκε στο κεφάλαιο (0), επομένως εδώ θα αναλύσουμε τη δική μας προσέγγιση για την υλοποίηση αυτής της στατιστικής αυτής μεθόδου εύρεσης όρων, χωρίς βέβαια να αλλοιώνουμε τη θεωρητική της βάση.

Η θεωρία, λοιπόν, συνοπτικά, είναι ότι η μετρική αυτή αξιολογεί ως υποψήφιους όρους πολυλεκτικές εκφράσεις χωρίς να περιορίζεται μόνο στο πλήθος των εμφανίσεων τους με την ίδια μορφή (απόλυτη συχνότητα), αλλά εντάσσοντας και ένα σκέλος υπολογισμού σχετικής συχνότητας των υπολέξεων τους σε άλλες εκφράσεις. Η σχετική συχνότητα επιδρά αρνητικά στο termhood ενός όρου. Η εξήγηση για αυτό είναι ότι οι όροι που εμφανίζονται σε πολλαπλές μεγαλύτερες εκφράσεις υποδηλώνουν ότι τα μέρη τους είναι ισχυρότερα. Άρα, για τις μεγαλύτερες εκφράσεις, το υπόλοιπο, δηλαδή η λέξη που δεν περιλαμβάνεται στον μικρότερο υποόρο, δεν είναι πιθανό να αποτελεί τεχνικό όρο, αφού εμφανίζεται με μικρότερη συχνότητα, δηλαδή είναι πιθανό να αποτελεί false addition στον κεντρικό όρο.

Εκτός από τον υπολογισμό του C-Value ως απλές μετρικές συχνότητας, χρειάζεται να υλοποιηθούν αναγκαία προηγούμενα στάδια, όπως ο εντοπισμός των υποψηφίων όρων με βάση μοτίβων POS-Tagging, η ομαδοποίηση και η έκφραση τους σε δομημένη μορφή και τελικά την αξιολόγηση και

υπολογισμό του termhood τους. Αναφερόμαστε δηλαδή στα παρακάτω βήματα, τα οποία η εφαρμογή μας αντιμετωπίζει με διακριτά υποπρογράμματα με το 1ο να τροφοδοτεί το 2ο και ούτω καθεξής:

1) Καταγραφή «Tokens» Δεδομένων και Εύρεση Υποψήφιων Όρων

2) Δημιουργία «Term Bag» και Εύρεση Παραλλαγών Όρων

3) Εύρεση Εμφωλευμένων Όρων και Υπολογισμός C-Value

Για την ανάπτυξη του αλγορίθμου, αξιοποιήθηκε η γλώσσα java. Θα αναφέρουμε ενδεικτικά κάποιες από τις σημαντικότερες βιβλιοθήκες και τεχνολογίες που χρησιμοποιήσαμε:

1) Stanford-postagger (MaxentTagger): Αυτή η βιβλιοθήκη χρησιμοποιήθηκε για την αναγνώριση του μέρους λόγου το οποίο αποτελούν οι λέξεις των κειμένων που χρησιμοποιούνται ως dataset. Δρα και παράγει tags για τις λέξεις με βάση τα σύμβολα του Penn-Treebank. [22]

2) Stanford-corenlp: Αυτή η βιβλιοθήκη χρησιμοποιήθηκε για την αναγνώριση των λέξεων από τα tokens που θα σταλούν στον MaxentTegger για προσδιορισμό του μέρους του λόγου το οποίο αποτελούν. Επιπλέον, μιας και η είσοδος είναι λίστες από τις επονομαζόμενες "HasWord" συμβολοσειρές, η βιβλιοθήκη αξιοποιήθηκε και για περεταίρω επαλήθευση ότι αποτελούν προτάσεις μέσα από την δομή "Sentence" που εμπεριείχε η βιβλιοθήκη.

3) edu.cmu.lti.ws4j: Αυτή η βιβλιοθήκη περιέχει τους αλγορίθμους του Wordnet Similarity For Java, που δίνει ένα API για σημασιολογική συσχέτιση και εύρεση ομοιοτήτων-συνωνύμων με βάση το WordNet[5] του Princeton. Χρησιμοποιήθηκε για εύρεση παραλλαγών όρων μέσω σημασιολογικής ομοιότητας (συνώνυμα).

4) jazzy-core: Αυτή η βιβλιοθήκη (Jazzy) περιέχει έναν spell-checker ο οποίος ελέγχει την ορθογραφία των λέξεων και δεδομένου ενός input file βρίσκει όρους με ελάχιστο πλήθος ορθογραφικών διαφορών με την λέξη που του εισαγουμε. Υλοποιεί μιας μορφής edit distance και χρησιμοποιήθηκε για την εύρεση παραλλαγών όρων λόγω ορθογραφικών διαφορών.

5) SQLite: Χρησιμοποιήσαμε μία ελαφριά σχεσιακή βάση δεδομένων προκειμένου να αποθηκεύουμε τα ενδιάμεσα στάδια πληροφορίας. Στα ακόλουθα σχήματα, τα σημεία όπου έχουμε συναλλαγές με τη βάση φαίνονται με τους κυλίνδρους που απεικονίζουν τους διάφορους πίνακες. Πέρα από την αξιοποίηση java tuples για να αποθηκεύσουμε ενδιάμεση συνδυαστική πληροφορία, σε αρκετά σημεία η δόμηση της πληροφορίας απαιτούσε πολλά attributes για κάθε υποψήφιο όρο, όπως το πλήθος των λέξεων που το αποτελούν, τα tokens που τον απαρτίζουν, η θέση του στα διάφορα κείμενα και τελικά

[5] Η online βιβλιοθήκη του wordnet: https://wordnet.princeton.edu/

τα στατιστικά στοιχεία για τις εμφανίσεις του στο κείμενο και το C-Value του. Με εξαίρεση τις σχετικές εμφανίσεις των tokens ενός term σε άλλους όρους εκτός από αυτόν, όπου χρησιμοποιήθηκε in-memory υλοποίηση, όλα τα υπόλοιπα υπάρχουν αποθηκευμένα σε SQLite Tables όπως οι term_phrase, term_nested, term_termhood, data_tokens. Λίγα λόγια για αυτούς τους πίνακες:

1. Ο data_tokens περιέχει σε πρωτογενή μορφή τα tokenized strings όλων των λέξεων του κειμένου (συνόλου προτάσεων) που έχει προκύψει μετά το preprocessing. Επίσης, περιέχει και τις stemmed μορφές τους, αλλά και τα POS tags τους, δηλαδή το Penn-Treebank [23] σύμβολο που αντιστοιχίζεται με το μέρος του λόγου που αποτελούν.

2. Ο term_phrase περιέχει τους υποψήφιους όρους, δηλαδή εκείνα τις ακολουθίες tokens από το data_tokens για τα οποία έχει βρεθεί ότι τα μέρη του λόγου τους (tags) σχηματίζουν ακολουθίες που είναι πιθανές ως υποψήφιοι όροι. Επίσης, έχει πληροφορίες για το πού βρίσκονται, όπως σε ποιο κείμενο, ποια γραμμή, τι μήκος έχουν πόσες λέξεις, καθώς επίσης και την κανονικοποιημένη μορφή τους μετά από το stemming αλλά και μετά από τη διαδικασία εύρεσης διαφοροποιήσεων όρων (expanded, βλέπε ενότητα 3.3.4)

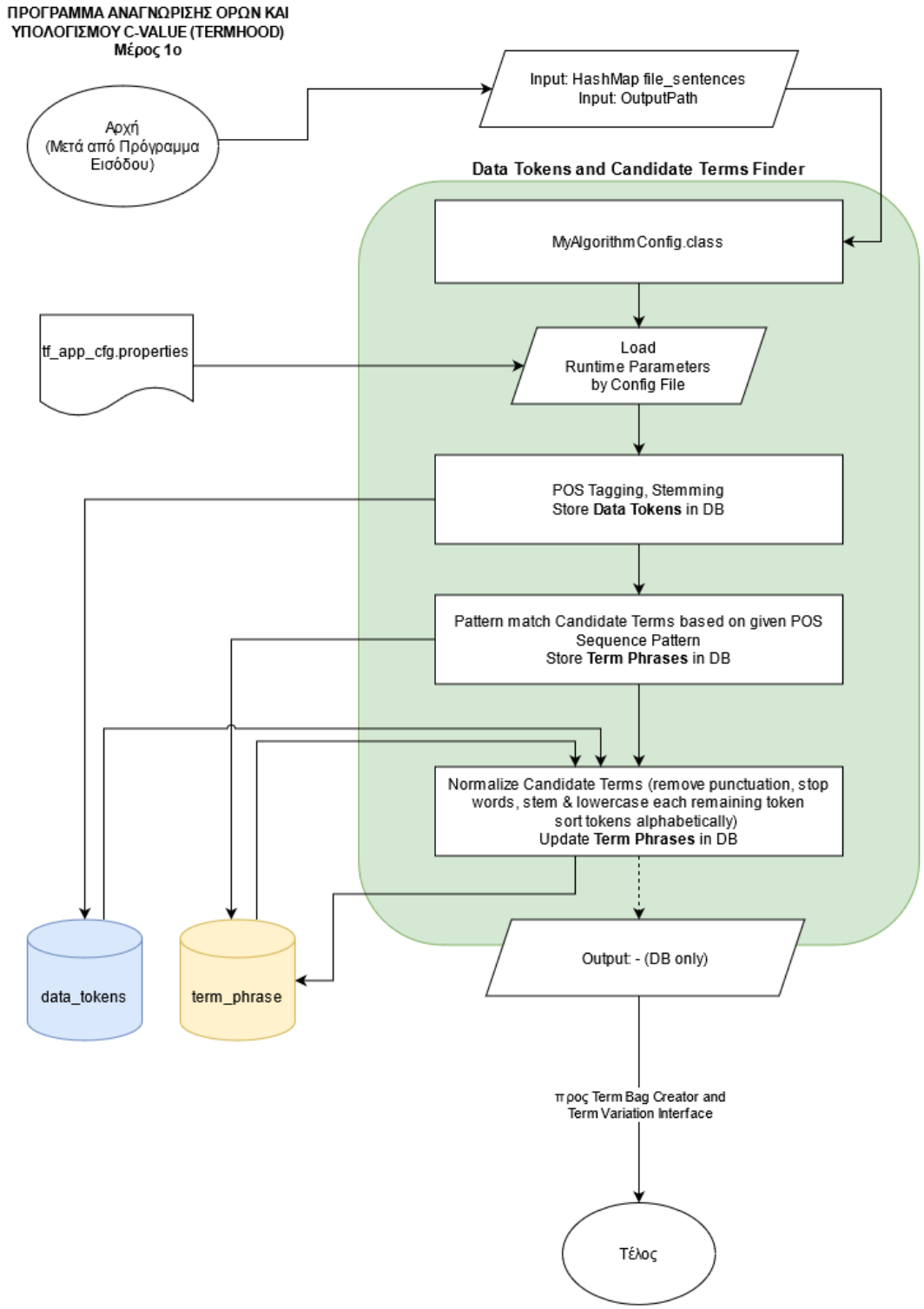
3. Ο term_nested είναι στην ουσία ένα mapping ανάμεσα στους όρους του term_phrase και εκφράζει τη συσχέτιση N-to-M (relation) parent-child για τους όρους. Ένας όρος είναι parent ενός δεύτερου όρου child, όταν όλα τα tokens του child εμπεριέχονται στον parent (δηλαδή κάθε child είναι υποσύνολο του ή των parents του)

4. Ο term_termhood περιέχει για κάθε term_phrase στατιστικά στοιχεία σε σχέση με αυτό και τις εμφανίσεις του στα κείμενα ή σε άλλους όρους. Έχει ως στήλες το c (από το c-value) και όλες τις παραμέτρους που χρειάζονται για τον υπολογισμό του c-value. Το περιεχόμενο του προέρχεται από τον term_phrase.

Κάθε ένα από τα υποπρογράμματα παράγει logs κατά την εκτέλεση του, μερικά εκ των οποίων θα παρουσιάσουμε με ένα παράδειγμα λειτουργίας ώστε να γίνει κατανοητή η χρησιμότητα του. Στόχευση της τμηματικής δομής του αλγορίθμου είναι η διευκόλυνση της παραμετροποίησης του για τα διαφορετικά πειράματα, δεδομένου ότι γίνεται πιο εύληπτη η πληροφορία για τα σημεία εισόδου και εξόδου.

Για τον λόγο αυτό, καθώς επίσης το πρόγραμμα είναι υπερβολικά περίπλοκο ώστε να μπορέσει η λειτουργία του να αναπαρασταθεί σχηματικά σε μία σελίδα, στις επόμενες ενότητες θα αναφερθούμε ξεχωριστά σε κάθε ένα από τα 3 στάδια διαδοχικά.

Αλγόριθμος Αναγνώρισης Όρων (Α.Α.Ο) Μέρος 1ο:



Σχήμα 6: Μέρος 1ο Αλγορίθμου Αναγνώρισης Όρων (Α.Α.Ο) – Data Tokens and Candidate Terms Finder

Στο πρώτο μέρος, σκοπός μας είναι από το απλό Mapping Αρχείων-Προτάσεων να καταλήξουμε στους υποψήφιους όρους.

Πρώτα κάνουμε την απαραίτητη προεργασία χρησιμοποιώντας το configuration file (*Αρχείο Διαμόρφωσης*) (για παράδειγμα, φόρτωση των stopwords). Επιλέγουμε τον POS-Tagger και τον Stemmer και για κάθε tokenized πρόταση διακρίνουμε τα επιμέρους της συστατικά-λέξεις (tokens) με απλό split ανά space (το tokenization έχει ήδη γίνει). Με τα tokens εντοπισμένα, είμαστε έτοιμοι να τα εισάγουμε στη SQLite βάση δεδομένων μας στον πίνακα data_tokens, που περιγράψαμε προηγουμένως.

Την ίδια στιγμή, τροφοδοτούμε στον POS Tagger την πρόταση ώστε να προσθέσει ως απόληξη στα tokens το μέρος του λόγου των λέξεων. Αυτήν την απόληξη αποθηκεύουμε ως ακολουθία και επιχειρούμε να βρούμε ποιες υπακολουθίες από αυτά κάνουν match στους σχηματισμούς του μοτίβου που έχουμε θέσει ως είσοδο. Π.χ. για τις λέξεις *«dry eyes»*, ο POS Tagger θα τις μετασχηματίσει σε «dry/JJ eyes/NN» και θα φτιάξουμε την ακολουθία JJ NN, η οποία θα πληροί το regexp μοτίβο (((NN|JJ))*NN). Επιπλέον ανάλυση για το POS Tagging ακολουθεί και στην ενότητα (5.2), όπου αναλύονται όλα τα μοτίβα που χρησιμοποιήσαμε στα πειράματα.

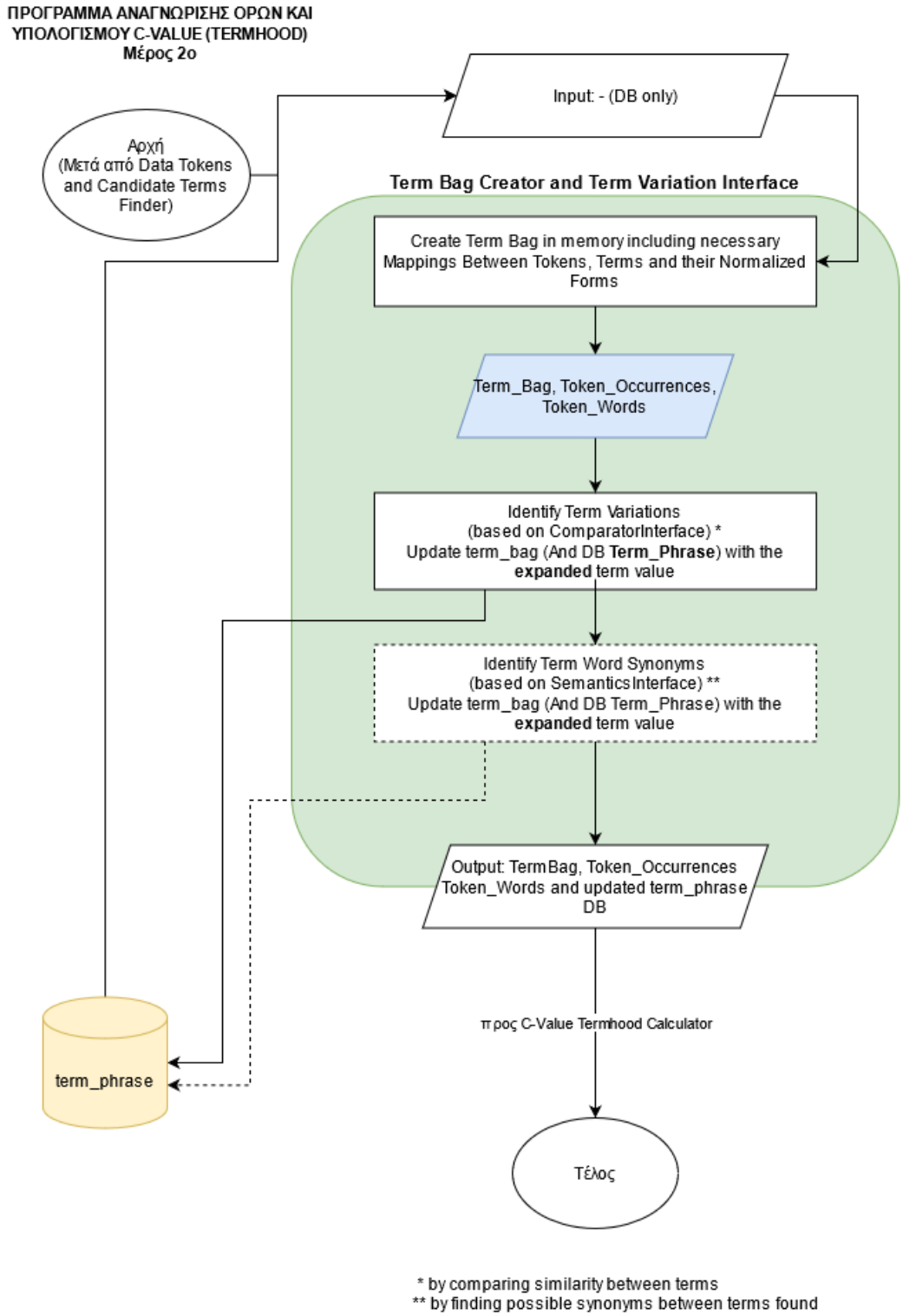
Αν η ακολουθία, όπως το dry eyes, πληροί το μοτίβο, την εισάγουμε στη βάση δεδομένων στον πίνακα term_phrase, με επιπλέον πληροφορίες για το πού βρέθηκε (αρχείο, γραμμή) και τι μήκος έχει. Αυτός ο πίνακας είναι ο βασικός πίνακας που θα χρησιμοποιήσουμε στη συνέχεια για να υπολογιστούν

Προβαίνουμε επίσης σε κανονικοποίηση του εντοπισμένου όρου, με εισαγωγή στο cand_terms των stems που αντιστοιχίζονται στα tokens που τον απαρτίζουν. Μάλιστα, αγνοούμε τα σημεία στίξης, εξαιρούμε από το stemming τα stopwords [24][6] και όσες λέξεις είναι κάτω των 3 χαρακτήρων. Ο λόγος που βάλαμε ελάχιστο μέγεθος λέξης ήταν για να μην συμπεριλάβουμε συντομεύσεις και ακρωνύμια στους όρους μας, για τα οποία θα χρειαζόταν πιο ενδελεχή ανάλυση. Αφού βρούμε τα stems, τα ταξινομούμε αλφαβητικά, κάτι πού θα μας βοηθήσει στο δεύτερο μέρος με τη δημιουργία του συνόλου «term-bag» που θα εξηγήσουμε σε λίγο.

Η έξοδος του 1ου μέρους έχει μηδενική επιστροφή, αλλά στην ουσία είναι οι 2 πίνακες στη βάση μας.

[6] Η λίστα που βάζουμε στις αναφορές ήταν η alpha version, την οποία εμπλουτίσαμε και με άλλες λέξεις που εμφανίζονταν στα iterations μας και manually διαπιστώναμε ότι αλλοίωναν την λειτουργικότητα του αλγορίθμου μας γιατί ενώ ήταν αρκετά συχνές, δεν παραπέμπουν νοηματικά σε ιατρικούς όρους.

Αλγόριθμος Αναγνώρισης Όρων (Α.Α.Ο) Μέρος 2ο:



Σχήμα 7: Μέρος 2ο Αλγορίθμου Αναγνώρισης Όρων (Α.Α.Ο) – Term Bag Creator and Term Variation Interface

Στο δεύτερο μέρος, με δεδομένους τους υποψήφιους όρους και την κανονικοποιημένη μορφή τους (stemmed), εκτελούμε το σκέλος του αλγορίθμου C-Value που ερμηνεύει τους κανόνες για τις πολυλεκτικές εκφράσεις. Στόχος μας είναι να ανιχνεύσουμε για κάθε πολυλεκτικό (αλλά και μονολεκτικό) υποψήφιο όρο σε ποιους άλλους όρους εμπεριέχεται κάθε token του.

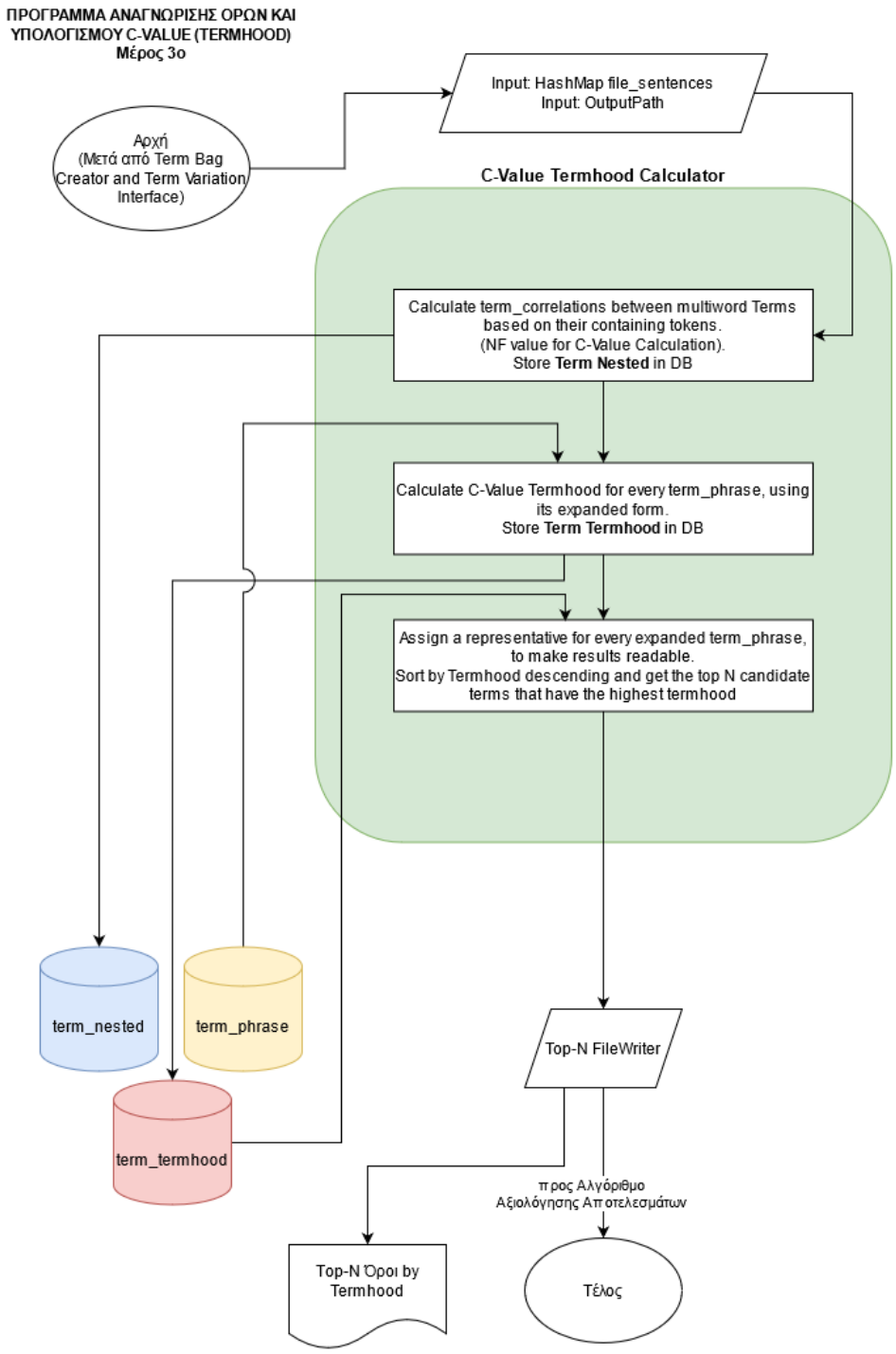
Για αυτό, χρειάζεται στην αρχή να δημιουργήσουμε μία δομή που θα έχει ως κλειδί τον κάθε όρο στο term_phrase (ή για την ακρίβεια το id του) και ως τιμή έναν πίνακα με τα tokens που τον απαρτίζουν. Μιας και η λογική μας είναι ότι δεν μας νοιάζει η σειρά εμφάνισης όρων, τα ταξινομούμε σε αύξουσα σειρά. Αυτό το hash_map το ονομάζουμε term_bag. Εκτός από το term_bag, δημιουργούμε άλλες 2 δομές: την token_occurrences η οποία εμπεριέχει ως κλειδί τα tokens και ως value μια λίστα με το πού αυτά εμφανίζονται στο κείμενο, ενώ η token_words αντιστοιχίζει τα tokens-stems με όλες τις δυνατές λέξεις που καταλήγουν μέσω stemming στο ίδιο stem.

Με αυτές τις 3 δομές, μπορούμε να προχωρήσουμε στην εύρεση variations και synonyms στους όρους μας. Από το term_bag, ξεκινάμε να αναζητούμε για κάθε όρο variation από τις λέξεις του συνόλου, συγκρίνοντας με όλα τα εμφανιζόμενα tokens στο κείμενο και επιλέγοντας αυτά που μορφολογικά διαφέρουν μέχρι 2 χαρακτήρες maximum. Όμοια και στη περίπτωση των synonyms, στην οποία κάνουμε χρήση του Wordnet Similarity For Java (ws4j), αν από ένα score και πάνω τα words ενός όρου έχουν συνώνυμα από words του κειμένου, επιλέγουμε κι αυτά. Με τους όμοιους μορφολογικά και σημασιολογικά χαρακτήρες επιλεγμένους, δημιουργούμε το επονομαζόμενο expanded column για τον υποψήφιο όρο. Πρόκειται στην ουσία για την ένωση συνόλων των συνώνυμων ή μορφολογικά όμοιων tokens σε ξεχωριστούς όρους. Έτσι, φτιάχνουμε στην ουσία ένα επεκταμένο normalized form για κάθε όρο αφού εκτός από τα stems των συστατικών λέξεων έχουμε και τα stems όλων των variations. Για τις ανάγκες της διπλωματικής αυτής εργασίας, αξιοποιήσαμε στα πειράματα μας μόνο το μορφολογικό σκέλος, καθώς για το σημασιολογικό σκέλος η βιβλιοθήκη του Wordnet[7] μας έβγαζε ελάχιστα αποτελέσματα, δεδομένου ότι είναι γενικής χρήσεως. Παραπέμπουμε στην ενότητα (3.3.4) για περεταίρω ανάλυση.

Αυτό το expanded το εισάγουμε ως column σε όλους τους όρους του term phrase που είχαν normalized υποσύνολό του και έτσι είμαστε έτοιμοι να προχωρήσουμε στον υπολογισμό του C-Value με βάση αυτό.

[7] Η βιβλιοθήκη του Wordnet: https://wordnet.princeton.edu/

Αλγόριθμος Αναγνώρισης Όρων (Α.Α.Ο) Μέρος 3ο:



Σχήμα 8: Μέρος 3ο Αλγορίθμου Αναγνώρισης Όρων (Α.Α.Ο) – C-Value Termhood Calculator

Στο 3ο μέρος γίνεται ο ουσιαστικός υπολογισμός του C-Value. Για αυτόν, εκτελούμε διαδοχικά βήματα που χτίζουν σιγά σιγά κάθε παράμετρο που απαιτείται για να υπολογιστεί το τελικό termhood.

Όπως αναφέραμε και νωρίτερα, για τον υπολογισμό του C-Value χρειαζόμαστε τη γνώση των εμφωλεύσεων των υποψήφιων όρων μας σε μεγαλύτερους υπερ-όρους. Έτσι, αξιοποιώντας τις δομές των token_occurrences και του term_bag που υπολογίσαμε στο προηγούμενο μέρος, έχουμε τη δυνατότητα για κάθε term_bag να ανιχνεύσουμε σε ποια άλλα terms υπάρχουν token_occurrences του. Αν όλα τα tokens ενός term_bag εμπεριέχονται σε αυτά ενός άλλου term_bag, τότε σημειώνουμε σε έναν νέο πίνακα που ονομάζουμε term_correlations τις σχέσεις παιδιού-πατέρα, για υπο-όρους που εντάσσονται σε υπερ-όρους. Αυτό καταλήγει να αποθηκευτεί στη βάση μας στον πίνακα term_nested.

Μετά από τη δημιουργία του term_nested, εισάγουμε από το term_phrase όλα τα expanded των υποψήφιων όρων στον τελικό μας πίνακα, term_termhood, μαζί με το πλήθος των όρων τους. Επειδή κάθε γραμμή στο term_phrase έχει ως primary key το cand_term, το πλήθος των όρων αφορά την αρχική μορφή του όρου.

Στη συνέχεια, υπολογίζουμε την απόλυτη συχνότητα εμφάνισης για κάθε expanded, για κάθε διαφορετικό πλήθος όρων. Αυτό θα είναι το *f(t)*. Ξεκινάμε τον υπολογισμό από τα expanded με το μεγαλύτερο πλήθος και προχωράμε προς τα επόμενα με το αμέσως επόμενα. Ο τύπος του C-Value είναι ο εξής [14],[13]:

$$C\text{-}value(t) = \begin{cases} \ln|t| \cdot f(t) & ,\text{if } S(t)=\varnothing \\ \ln|t| \cdot \left(f(t) - \frac{1}{|S(t)|}\sum_{s \in S(t)} f(s)\right) & ,\text{if } S(t)\neq\varnothing \end{cases}$$

Όπου,

- *t* το expanded version κάθε όρου μας,
- *S(t)* το σύνολο των όρων που εμπεριέχουν το t (υπερόροι),
- *f(t)* η απόλυτη συχνότητα του όρου μας,
- *f(s)* η απόλυτη συχνότητα των υπερόρων μας (και θέλουμε το άθροισμα όλων)
- |*S(t)*| το πλήθος των διαφορετικών υπερόρων του όρου μας.

Άρα, πρέπει να υπολογίσουμε τα *f(s)* και |*S(t)*| Εδώ αξιοποιούμε το γεγονός ότι έχουμε υπολογίσει το term_nested νωρίτερα και εντοπίζουμε με ευκολία το |*S(t)*| ως ένα απλό count των πατεράδων για το

συγκεκριμένο παιδί. Το ίδιο και για το πλήθος, απλά σε αυτή την περίπτωση χρειάζεται να ανατρέξουμε στο term_phrase.

Αφού υπολογιστούν οι άνωθι μεταβλητές του τύπου μας, υπολογίζουμε το C-Value για κάθε expanded μορφή των όρων μας. Όμως, για να έχουμε παρουσιάσιμα αποτελέσματα, δεν αρκεί να δείξουμε ένα σύνολο από stems που έχουν παραχθεί από την επέκταση πολλών normalized όρων μεταξύ τους. Για αυτό το λόγο, υπολογίζουμε έναν representative για κάθε expanded όρο, ανατρέχοντας στην αρχική αναγνώσιμη μορφή με την οποία είχαμε εντοπίσει τον όρο. Επιλέγουμε αυτόν τον όρο με τις περισσότερες εμφανίσεις από όλους αυτούς με τη κοινή ρίζα-expanded.

Μετά από αυτό το βήμα, έχουμε συγκεντρώσει στη βάση δεδομένων μας και συγκεκριμένα στον πίνακα term_termhood όλη την πληροφορία για το termhood των υποψήφιων όρων μας. Αρκεί λοιπόν να εξάγουμε τα δεδομένα μας σε έναν συγκεκριμένο output folder με βάση το Αρχείο Διαμόρφωσης, επιλέγοντας τους Top-N όρους, μαζί με το C-Value και την normalized-stemmed εκδοχή τους, ώστε να τους τροφοδοτήσουμε στον αλγόριθμο αξιολόγησης αποτελεσμάτων.

3.3.4 Σύγκριση Συμβολοακολουθιών – Διαφοροποιήσεις Όρων

Οι όροι που παρουσιάζονται στα σύγχρονα τεχνικά κείμενα πολύ συχνά διαθέτουν παραλλαγές γεγονός που καθιστά δυσκολότερη την ανεύρεση τους από την πρώτη μορφή του συστήματος μας. Ταυτόχρονα, η ίδια η αγγλική γλώσσα παρουσιάζει από μόνη της ποικιλία γραμματικών κανόνων, όπως ο ενικός και ο πληθυντικός αριθμός, αλλά και εκτενές λεξιλόγια, με αρκετά συνώνυμα για μεγάλο μέρος από τα ουσιαστικά που αποτελούν την βάση των όρων.

Αυτό οδήγησε στην σταδιακή εξέλιξη των τεχνικών που χρησιμοποιούσαμε για την αντιστοίχιση φράσεων οι οποίες συνιστούν παραλλαγές η μία της άλλης. Αυτή τη σταδιακή βελτίωση της προσέγγισης μας θα περιγράψουμε παρακάτω.

Αρχικά, ο πιο στοιχειώδης τρόπος αντιμετώπισης του προβλήματος ήταν η απλή αντιστοίχιση των συμβολοσειρών για ταυτοσημία. Αναζητώντας τις λέξεις που αποτελούσαν τους υποψήφιους όρους μας, και μάλιστα διατηρώντας τη σειρά σε πολυλεκτικούς όρους, συνυπολογίζαμε στις συχνότητες με κριτήριο την απόλυτη ταύτιση όλων των λέξεων κάθε όρου στη σωστή σειρά. Αυτό αποτελεί μια απλοϊκή προσέγγιση η οποία δεν δίνει περιθώριο για καμία διαφοροποίηση στους όρους. Με την υλοποίηση αυτή, απορρίπταμε κάθε διαφοροποίηση στη μορφή μιας λέξης, με αποτέλεσμα πληθυντικοί αριθμοί, συνώνυμα και ορθογραφικά λάθη/παραλλαγές να οδηγούσαν στην εμφάνιση πολλών εγγραφών για κάθε όρο, και έτσι στην λαθεμένη αποτίμηση της termhood λόγω του μη αναγκαίου διαμοιρασμού της.

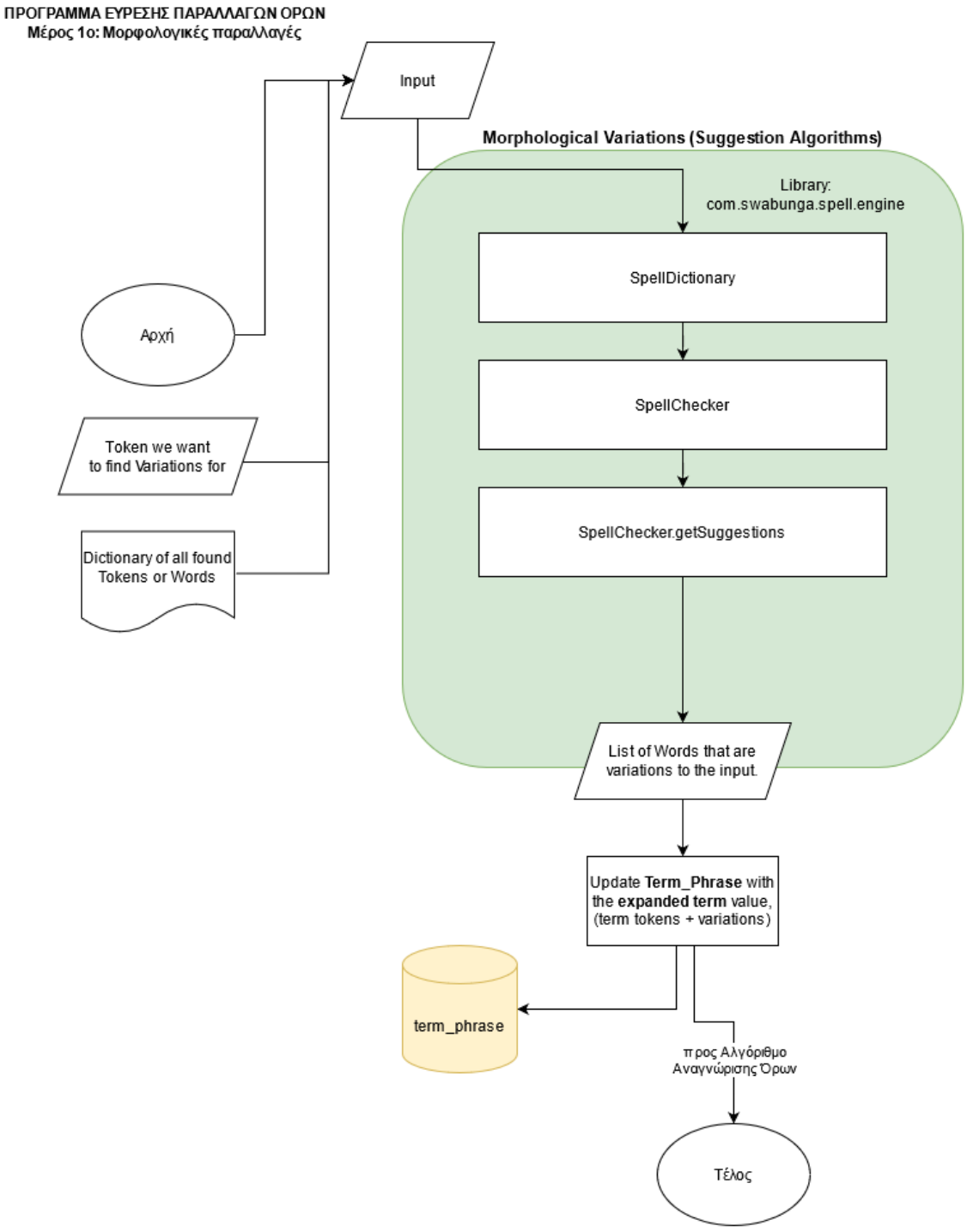
Για να επεκτείνουμε αυτό το μοντέλο, σύντομα συνειδητοποιήσαμε ότι έπρεπε να ακολουθηθεί η τεχνική εύρεσης της ρίζας κάθε λέξης, έτσι ώστε τα προαναφερθέντα προβλήματα με τις διαφορετικές παραλλαγές κάθε λέξης λόγω γραμματικής και ορθογραφικής μορφολογίας να προσπεραστούν μέσα από την σύγκριση των ριζών των υποψηφίων όρων και όχι της αρχικής τους μορφής. Μιλούμε επομένως για τη διαδικασία εύρεσης της ρίζας μίας λέξης δηλαδή του stemming. Όπως έχει διαπιστωθεί σε πολλαπλές μελέτες επί της αναγνώρισης όρων, οι αλγόριθμοι που υλοποιούν stemming χρησιμεύουν στην κανονικοποίηση των λεκτικών οντοτήτων με υπαγωγή στην ίδια ρίζα όρων με γλωσσική ή σημασιολογική συγγένεια. Το αποτέλεσμα είναι να λύνονται προβλήματα ορθογραφικής και μορφολογικής παραλλαγής των όρων με γρήγορη.

Στη συνέχεια, ενισχύσαμε την χαλαρότητα στο κριτήριο, επιτρέποντας fuzziness. Το fuzziness είναι στην ουσία η διαφορά μεταξύ λέξεων αρχικά θεωρούμενων ξεχωριστών όρων, η οποία τελικά ορίζεται αποδεκτή μέχρι ένα μικρό πλήθος χαρακτήρων για να θεωρηθούν οι 2 αυτοί όροι ίδιοι. Έτσι, μπορούμε να αντιμετωπίσουμε το πρόβλημα του αριθμού στα αγγλικά ουσιαστικά. Την ίδια στιγμή, όμως, διατρέχουμε αυξημένο κίνδυνο false positives, δηλαδή να αντιστοιχίσουμε δύο όρους με σχεδόν ίδια ορθογραφία αλλά εντελώς διαφορετικό νόημα. Η ισορροπία αυτή προδιαθέτει την ιδέα ότι θα χρειαστούμε πολλαπλές επαναλήψεις, επομένως αποτελεί ένα εφαλτήριο για να χρησιμοποιήσουμε σε μελλοντικές εκδόσεις του συστήματος τεχνικές μηχανικής μάθησης ως μεθόδους επαναλήψεων και επανεκπαίδευσης τους συστήματος.

Στον αλγόριθμο μας, οι διαφοροποιήσεις-παραλλαγές υλοποιούνται με διακριτά interfaces τα οποία επιτρέπουν την επιλογή διαφορετικών υλοποιήσεων αλγορίθμων για την ίδια διεργασία, καθώς και την απενεργοποίηση των λειτουργιών για την εκτέλεση συγκριτικών πειραμάτων ώστε να βρεθεί η επίδραση κάθε λειτουργίας ξεχωριστά. Η τμηματοποίηση του προγράμματος παρέχει μεγαλύτερο έλεγχο στις παραμέτρους που θέλουμε να εξετάσουμε και επιτρέπει ταχύτερη ανατροφοδότηση του συστήματος σε συνάρτηση με τα αποτελέσματα που βγάζουμε.

Έτσι, διακρίνουμε τα εξής με βάση το παρακάτω 2 σχηματικά διαγράμματα:

Πρόγραμμα Εύρεσης Παραλλαγών Όρων: Μορφολογικές Παραλλαγές



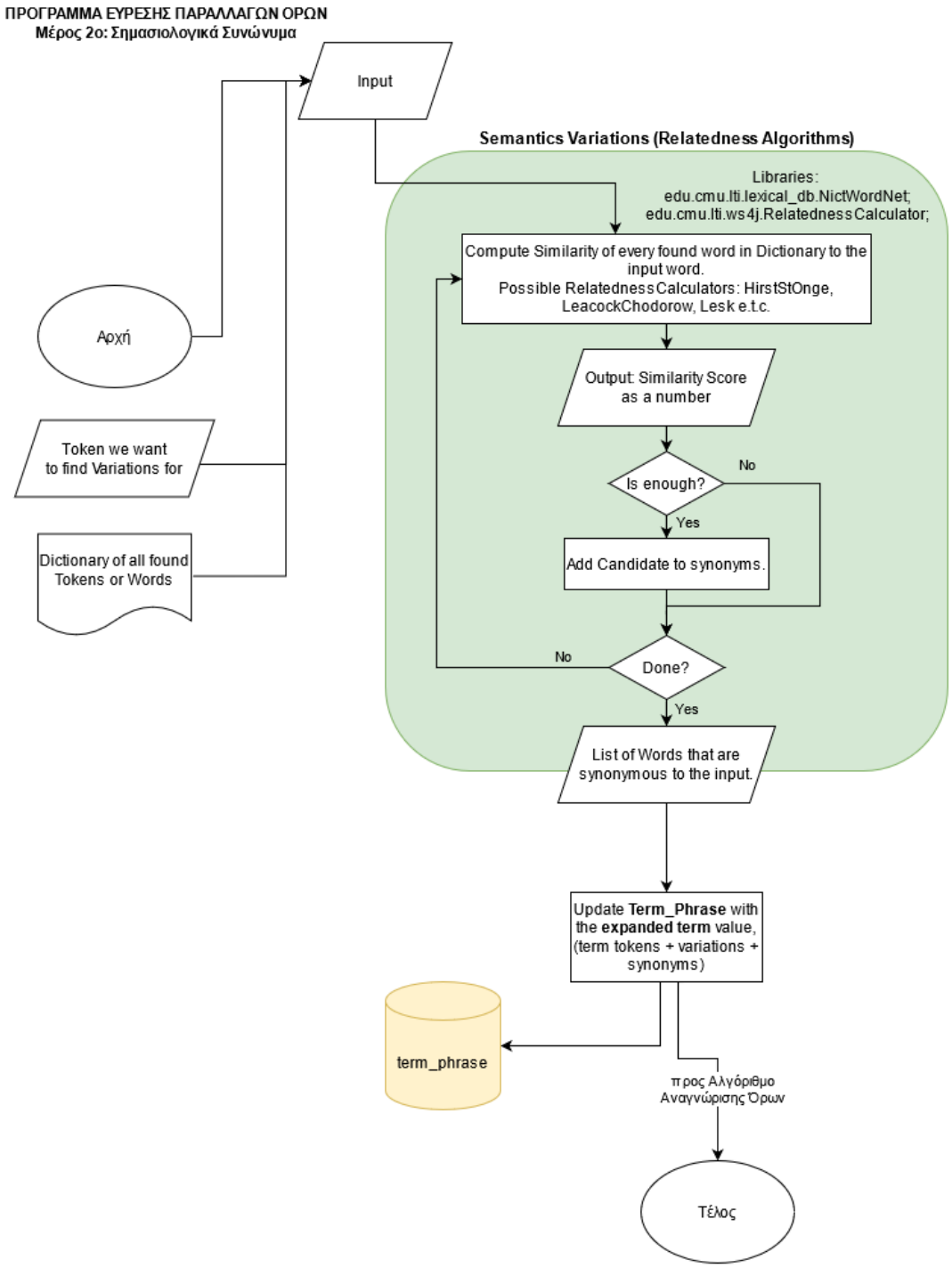
Σχήμα 9: Διεπαφή Μορφολογικών Παραλλαγών

Στην περίπτωση των μορφολογικών παραλλαγών, κάνουμε χρήση της βιβλιοθήκης jazzy για SpellChecking. Για κάθε λέξη όρου, βρίσκουμε παραλλαγές της με αναζήτηση σε λεξικό που δημιουργήσαμε με τη βοήθεια της δομής token_word που εμπεριέχει τις αρχικές λέξεις. Από το token_word, βρίσκουμε αντιστοιχία με κάθε token_occurrence. Για κάθε διακριτό token_occurrence, το current_token_occurrence, καλούμε SpellChecking με dictionary που υπολογίσαμε μία φορά στην αρχή. Αν υπάρχει διαφορά μέχρι το πολύ 2 χαρακτήρες (όμοια με έναν αλγόριθμο edit distance) επιλέγονται ως suggestions. Με αυτά τα suggestions εμπλουτίζουμε το term_bag για αυτόν τον όρο. Μόλις ολοκληρωθεί η διαδικασία για κάθε token όρου, τα νέα term_bags με τις παραλλαγές των tokens κάνουν update τη στήλη expanded των αντίστοιχων υποψήφιων όρων στον πίνακα term_phrase.

Πρόγραμμα Εύρεσης Παραλλαγών Όρων: Σημασιολογικές Παραλλαγές

Στην περίπτωση των σημασιολογικών παραλλαγών, ο μηχανισμός είναι παρόμοιος με πριν. Αντί για spellchecker χρησιμοποιούμε έναν Relatedness Algorithm ο οποίος πάλι για κάθε λέξη επιχειρεί να βρει συνώνυμα εντός του κειμένου με γνωστούς relatedness algorithms όπως ο HirstStOnge, LeacockChodorow και άλλοι (παραπέμπουμε στο εξής paper [25] για περεταίρω ανάλυση του μηχανισμού που δουλεύουν, μιας και η πολυπλοκότητα τους δεν μπορεί να αναλυθεί στα πλαίσια αυτής της εργασίας) και να εμπλουτίσει την στήλη expanded του όρου που τα περιέχει. Έτσι όροι που έχουν όλες τις λέξεις τους συνώνυμα καταλήγουν να έχουν το ίδιο expanded και τις κατηγοριοποιεί μαζί ο αλγόριθμος C-Value (το οποίο αναγνωρίζει και αν έχουν σχέση πατέρα-παιδιού καθώς όπως αναφέραμε στην προηγούμενη ενότητα, ο αλγόριθμος λαμβάνει και αυτές τις σχέσεις υπόψιν).

Για τις ανάγκες της σημασιολογικής αναζήτησης συνωνύμων, χρησιμοποιήσαμε βιβλιοθήκες που αξιοποιούν το λεξικό Wordnet του Princeton. Δυστυχώς όμως, βρήκαμε περιορισμένη βελτίωση των αποτελεσμάτων με χρήση του λεξικού αυτού (σε κάποιες περιπτώσεις μικρών κειμένων, δεν έβρισκε συνώνυμα, πολύ περισσότερο για ιατρικούς όρους που ούτε καν τους περιείχε), επομένως δεν τον συμπεριλάβαμε στα πειράματα μας. Παρ'όλα αυτά η υλοποίηση του υπάρχει ως απενεργοποιημένο class του Semantics Interface. Ακολουθεί το αντίστοιχο σχηματικό διάγραμμα:



Σχήμα 10: Διεπαφή Σημασιολογικών Παραλλαγών

4 Δοκιμή και Αξιολόγηση Συστήματος

Με δεδομένη τη δομή και τα συστατικά μέρη του αλγορίθμου, χρειάστηκε να γίνει αντιμετώπιση πιθανών προβλημάτων αλλά και να δοκιμαστεί η αποτελεσματικότητα του συστήματος μας. Η λειτουργία και η αρχιτεκτονική του Συστήματος Δοκιμής και Αξιολόγησης είναι αρκετά απλή και την περιγράφουμε στο παράρτημα αναλυτικότερα (7.1). Εν συντομία, ακολουθεί αντίστοιχες αρχές parsing και preprocessing με τον κύριο αλγόριθμο, αλλά στη συνέχεια τροφοδοτείται από τις εξόδους του κύριου αλγορίθμου και από αρχεία με «γνωστούς» όρους. Με βάση τους «γνωστούς όρους» και το πόσους από αυτούς θα εντοπίσει ότι ο αλγόριθμος μας έχει βρει, αξιολογεί την εφαρμογή μας, βρίσκει «άγνωστους όρους» που ο αλγόριθμος μας θεωρεί πιθανόν να αποτελούν όρους και μας δίνει μετρικές σχετικά με τα υπολογιζόμενα C-Values.

Για τους σκοπούς της αξιολόγησης, χρειάστηκε να αναζητήσουμε σύντομα test αρχεία εισόδου για τα οποία γνωρίζουμε εκ των προτέρων τι όρους θέλουμε να εντοπίσουμε. Ένα από αυτά τα «γνωστά» κείμενα εισόδου ήταν η εγγραφή της αγγλικής Wikipedia για το σύνδρομο Sjogren[8] από την οποία εξάγαμε τις πρώτες 4 εισαγωγικές παραγράφους. Η διαδικασία αξιολόγησης που ακολουθήσαμε ήταν η εξής: σε πρώτη φάση, γράψαμε manually τους όρους που βρήκαμε από μόνοι μας στο κείμενο εισόδου. Στη συγκεκριμένη περίπτωση εντοπίσαμε τους ακόλουθους 32 όρους:

Sjögren's syndrome	*long-term autoimmune disease*	*body's moisture-producing glands*	*organs systems*	*lungs*
kidneys	*nervous system*	*dry mouth*	*dry eyes*	*pain*
fatigue	*dry skin*	*symptoms*	*chronic cough*	*joint pains*
vaginal dryness	*other health problems*	*connective tissue disorder*	*rheumatoid arthritis*	*autoimmune diseases*
inflammation	*diagnosis*	*surgery*	*muscle pain*	*ibuprofen*
antihistamines	*lip biopsy*	*diagnostic test*	*disease*	*autoimmune disorder*
life expectancy	*muscle*			

Πίνακας 1: Εντοπισμένοι από εμάς ιατρικοί όροι στην εισαγωγή της Wikipedia για το σύνδρομο Sjogren

[8] Εγγραφή της Wikipedia για το Sjogren Syndrome: https://en.wikipedia.org/wiki/Sj%C3%B6gren_syndrome

Είναι εμφανές ότι σε αυτή τη διαδικασία, ιδανικό θα ήταν να υπάρχει επέμβαση από κάποιον ειδικό στον κλάδο της ιατρικής, αλλά για της ανάγκες της διπλωματικής αυτό δεν χρειάστηκε.

Στη συνέχεια, ανεξάρτητα από τους όρους αυτούς, βάζουμε τον αλγόριθμο μας να δουλέψει με είσοδο το μοναδικό αυτό άρθρο της Wikipedia με τις 4 παραγράφους. Τέλος, αφού υπολογίσει το C-Value για τους υποψήφιους όρους που βρήκε, τους εξάγουμε με φθίνουσα τιμή C-Value και εκτιμούμε τι ποσοστό όρων βρήκαμε, πόσο ψηλά τοποθετήθηκαν σε σύγκριση με άλλους προτεινόμενους όρους, ποιοι όροι χάθηκαν και γιατί.

4.1 Συντακτικά Μοτίβα

Για να δοκιμάσουμε τη λειτουργία του αλγορίθμου μας, χρειάστηκε να έχουμε έτοιμες τουλάχιστον τις βασικότερες παραμετροποιήσεις. Στην περίπτωση μας, ιδιαίτερη σημασία έχει η επιλογή των συντακτικών μοτίβων με βάση των οποίων γίνεται η ανίχνευση των υποψηφίων όρων. Για να διαπιστώσουμε ότι ο αλγόριθμος μας εντοπίζει υποψήφιους όρους ανάλογα με το μοτίβο που επιλέγουμε, προετοιμάσαμε με τη βοήθεια του Προγράμματος Διαχείρισης Υποσυστημάτων 2 διαφορετικές πειραματικές εκτελέσεις. Συγκεκριμένα, εκτελέσαμε την αξιολόγηση 2 φορές για 2 παραλλαγές μοτίβων POS-Tagging (Patterns) τα οποία είναι τα εξής:

1. (((NN|JJ)(| HYPH))*NN)
2. (((((NN|JJ))*NN) IN (((NN|JJ))*NN))|
   ((NN|JJ)*NN POS (NN|JJ)*NN))|
   (((NN|JJ)(| HYPH))*NN)

Η 1η θα αναφέρεται στο εξής ως Simple Pattern ενώ η 2η ως Complex Pattern. Ίσως ο καλύτερος και συντομότερος τρόπος να τις εξηγήσουμε είναι με παραδείγματα:

Στην περίπτωση του Simple Pattern θέλουμε μία ακολουθία από ουσιαστικά ή επίθετα, όσα θέλουμε (ακόμα και κανένα), αρκεί να διαχωρίζονται από κενό ( ) ή κάποιο ενωτικό (-) και να καταλήγουν στη συνέχεια σε ένα ουσιαστικό. Π.χ. η φράση *«Hemolytic Anemia»* αποτελείται από 1 επίθετο (*Hemolytic*) και λήγει με ένα ουσιαστικό *(JN - valid)*, άρα είναι έγκυρη για το Simple Pattern. Όμοια και το «*Anemia*» από μόνο του θα γίνει match ως 1 ουσιαστικό *(N - valid)*, όμως το «*Hemolytic*» από μόνο του δεν γίνεται match, αφού αρχίζει μεν ορθά με 1 επίθετο αλλά δε λήγει σε κανένα ουσιαστικό *(J – invalid)*.

Η 2η την οποία είπαμε ότι θα ονομάζουμε Complex Pattern είναι υπερσύνολο της πρώτης, επομένως κάθετι valid στην 1η είναι valid και εδώ όμως περιλαμβάνει και περιπτώσεις και μορφολογίες όρων που

δεν είναι valid στην 1η. Συγκεκριμένα, δέχεται επιπλέον περιπτώσεις όπως το εξής: «*Anemia of Chronic Disease*» δηλαδή συνθέσεις με 2 όρους από Simple Pattern που διαχωρίζονται από προθέσεις (όπως εδώ με το «*of*» ανάμεσα στα valid Simple Patterns των «*Anemia*» και «*Chronic Disease*»). Όμοια, δέχεται με τον ίδιο μηχανισμό και περιπτώσεις με 2 όρους από Simple Pattern που διαχωρίζονται από το κτητικό « 's». Άρα, ενώ για παράδειγμα το Simple Pattern δε θα δεχόταν το «*Sjogren's Syndrome*», το Complex το δέχεται.

Αφού, λοιπόν, αναλύσαμε με ποια μοτίβα θα εκτελεστούν τα δοκιμαστικά πειράματα, ας περάσουμε στην αξιολόγηση των ίδιων των πειραμάτων.

4.2 Αυτόματος Εντοπισμός Όρων

Αρχικά, παρουσιάζονται τα αποτελέσματα σχετικά με το πλήθος των όρων που βρέθηκαν καθώς και επιπλέον μετρικές σχετικά με την ακρίβεια και την ανάκληση του αλγορίθμου. Στόχος μας είναι να μετρήσουμε στο μέτρο του δυνατού σε ποιον βαθμό η αυτόματη αναγνώριση όρων είναι ισχυρότερη από την ανθρώπινη με το να εκτιμήσουμε την ποσότητα και (όσο είναι δυνατόν, ως μη ειδικοί) την ποιότητα των όρων που βρέθηκαν.

Η γενική εικόνα που αποκομίσαμε είναι ότι το σύστημα εντοπίζει πολύ περισσότερους όρους από ό,τι το ανθρώπινο μάτι. Αναφερόμενοι πάλι στις μετρικές της ακρίβειας *(precision)* και ανάκλησης *(recall),* θα μπορούσαμε να πούμε ότι έχουμε υψηλή ανάκληση αλλά η ακρίβεια είναι χαμηλή, όχι τόσο γιατί ευθύνεται το σύστημα αυτό καθαυτό αλλά περισσότερο γιατί οι όροι που εμείς εντοπίσαμε είναι λίγοι και κάποιοι άγνωστοι όροι είναι πράγματι ιατρικοί, ή γιατί το κείμενο δεν είναι αμιγώς επιστημονικό και οι άγνωστοι όροι δεν είναι ιατρικοί.

Κάνουμε την υπόθεση ότι στην Wikipedia οι όροι που θα βρεθούν θα είναι ιατρικοί και περισσότερο υστερούμε εμείς ως άνθρωποι να τους βρούμε, παρά ότι στη Wikipedia υπάρχουν μη σχετικοί όροι. Η υπόθεση θα επαληθευτεί αν δούμε ποιοτικά σε τι βαθμό είναι πράγματι υποψήφιοι όροι, περίπτωση στην οποία θα δείξουμε ότι η εφαρμογή μας είναι αποδοτική. Ορίζουμε τον τύπο της ανάκλησης με βάση όσα αναφέραμε στην ενότητα (2.5) και βασίζονται στη δημοσίευση [19] :

$$ανάκληση = \frac{Γνωστοί\ όροι\ που\ βρέθηκαν}{Γνωστοί\ Όροι\ που\ βρέθηκαν\ +\ Γνωστοί\ Όροι\ που\ δεν\ βρέθηκαν}$$

Τύπος 1: ανάκληση

Όμοια, ορίζουμε και τον τύπο της ακρίβειας με βάση πάλι όσα αναφέραμε στην ενότητα (2.5) και στη δημοσίευση [19] :

$$ακρίβεια\ = \frac{Γνωστοί\ όροι\ που\ βρέθηκαν}{Γνωστοί\ Όροι\ που\ βρέθηκαν\ +\ Άγνωστοί\ Όροι\ που\ βρέθηκαν}$$

Τύπος 2: ακρίβεια

Όμως, δεν έχουμε συμπεριλάβει καθόλου στον τύπο της ακρίβειας την επαλήθευση των άγνωστων όρων, οι οποίοι πρέπει να ληφθούν ως true positives για την τελική ακρίβεια. Θα τους ονομάσουμε καταχρηστικά «Αγνώστους Ιατρικούς Όρους» (δεν είμαστε γιατροί για να το αποφασίσουμε εμείς αλλά χάριν της μελέτης προχωρούμε με αυτόν τον τρόπο). Ορίζουμε δηλαδή την τελική ακρίβεια ως:

$$τελική\ ακρίβεια\ = \frac{Γνωστοί\ όροι\ που\ βρέθηκαν\ +\ Άγνωστοι\ "Ιατρικοί"\ Όροι}{Γνωστοί\ Όροι\ που\ βρέθηκαν\ +\ Άγνωστοί\ Όροι\ που\ βρέθηκαν}$$

Τύπος 3: τελική ακρίβεια

Στην ουσία, επεκτείνουμε το μοντέλο μας «γνωστών» όρων ώστε να περιέχει και τους άγνωστους όρους που κρίνουμε ως Ιατρικούς. Στο μέλλον, αυτή η «κριτική επιλογή» πρέπει να γίνεται από κάποιον ιατρικό ειδικό.

Για να γίνει, όμως, κατανοητή η μελέτη μας, σε ένα πρώτο στάδιο θα εστιάσουμε ποσοτικά στα αποτελέσματα που βγάλαμε, ενώ θα ακολουθήσει στη συνέχεια μια ποιοτική ανάλυση με αναφορά σε συγκεκριμένους όρους κάθε κατηγορίας και με πίνακες με αυτούς.

Ο παρακάτω πίνακας μας δείχνει τί ποσοστό ανάκλησης πετύχαμε ως προς τους όρους που βρήκαμε εμείς οι ίδιοι διαβάζοντας το κείμενο.

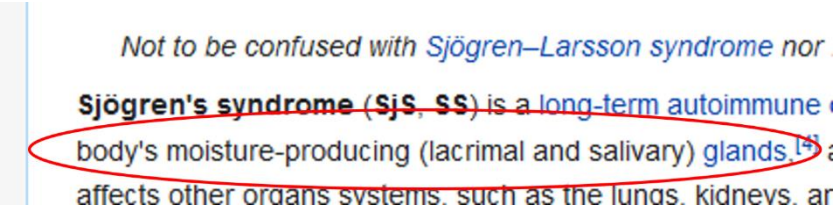
TermFinder (Ανάκληση) Προτεινόμενοι Όροι	**Simple Pattern**	**Complex Pattern**
Γνωστοί Όροι που βρέθηκαν	29	30
Γνωστοί Όροι που δεν βρέθηκαν	3	2
Όλοι οι Γνωστοί Όροι (Βρέθηκαν + Δεν βρέθηκαν)	32	32
Ανάκληση	90.63%	93.75%

Πίνακας 2: Αξιολόγηση Ανάκλησης Συστήματος – Δοκιμαστική Εκτέλεση

Από τα παραπάνω φαίνεται ότι χάνουμε 3 όρους στην περίπτωση του απλού μοτίβου και 2 στην περίπτωση του σύνθετου. Οι απώλειες μας είναι λογικές για κάθε περίπτωση για διαφορετικούς λόγους, ενώ δεδομένου ότι έχουμε ήδη δείξει ποιους όρους αναμένουμε να βρούμε και σχεδόν όλους τους βρήκαμε, θα ασχοληθούμε μόνο με όσους χάσαμε. Αυτοί είναι οι εξής:

Not Found Known Terms (Simple Pattern)		**Not Found Known Terms (Complex Pattern)**	
body's moisture-producing glands	*diagnostic test*	*body's moisture-producing glands*	*diagnostic test*
Sjögren's syndrome			

Ξεκινώντας με τα απλά, το "*Sjögren's syndrome"* χάνεται στην περίπτωση του απλού μοτίβου γιατί απλά δεν προβλέπει να κάνει handle το κτητικό "*'s"*. Για το "*body's moisture-producing glands"* βλέπουμε ότι στο κείμενο διακόπτει τον όρο με παρενθέσεις. Ένας άνθρωπος μπορεί να το διακρίνει, αλλά εδώ θα έπρεπε να κάνουμε το μοτίβο μας αρκετά πιο σύνθετο (δεν αρκεί να αγνοήσουμε απλά τις παρενθέσεις, γιατί αυτή η λύση μας απορρίπτει όρους που πιθανώς να εμπεριέχονται εξ 'ολοκλήρου εντός των παρενθέσεων – θα έπρεπε να κοιτάμε μοτίβα που εμπεριέχουν παρενθέσεις αλλά να αγνοούμε το περιεχόμενο τους όταν χτίζουμε την ακολουθία για τους όρους αριστερά και δεξιά από την περίπτωση – πράγμα αρκετά πιο πολύπλοκο από το συντακτικό μοτίβο που περιγράψαμε)



Εικόνα 1: Screenshot από Wikipedia στο σημείο που εμφανίζεται (μόνο 1 φορά) ο όρος που ανιχνεύσαμε εμείς αλλά όχι ο TermFinder

Για το *"diagnostic test"*, παρατηρούμε ότι στους άγνωστους όρους υπάρχει με τις μορφές *"specific extant diagnostic test"* και "*non-invasive diagnostic tests"* ενώ δεν υπάρχει αναφορά του αυτοτελώς. Με δεδομένο ότι ο c-value θεωρεί θετική την εμφάνιση μεγαλύτερων όρων (αν δεν εμφανίζονται πολλές φορές αυτοτελώς οι υποόροι τους) είναι λογικό να μην επιλέγει τον σχηματισμό diagnostic test από μόνο του. Γενικά, το ποσοστό ανάκλησης φαίνεται να είναι αρκετά ψηλά και στις 2 περιπτώσεις συντακτικών μοτίβων, άρα εντοπίζει ως όρους αυτούς που θεωρούμε γνωστούς εμείς. Ο επόμενος πίνακας που αφορά την ακρίβεια του αλγορίθμου μας, προκύπτει θεωρώντας επιτυχημένους μόνο τους 29 (και 30) όρους που είναι «γνωστοί». Σημειώνουμε ότι αυτή η ακρίβεια θα αναθεωρηθεί προς τα πάνω όταν συμπεριλάβουμε και τους άγνωστους «ιατρικούς όρους»:

TermFinder (Ακρίβεια) Προτεινόμενοι Όροι	**Simple Pattern**	**Complex Pattern**
Γνωστοί Όροι που βρέθηκαν	29	30
Άγνωστοι Όροι που βρέθηκαν	62	58
Όλοι οι Όροι που βρέθηκαν (Γνωστοί + Άγνωστοι)	91	87
Ακρίβεια	31,87%	34,48%

Πίνακας 3: Αξιολόγηση Ακρίβειας Συστήματος – Δοκιμαστική Εκτέλεση

Με τον πρώτο ορισμό της ακρίβειας, true positive θεωρούμε τους 31 γνωστούς όρους και μόνο (από τους οποίους στο simple βρήκαμε 29). Για να αξιολογήσουμε σωστά τον αλγόριθμο μας, από τους 62 προτεινόμενους όρους, αξιολογήσαμε 41 ως «Ιατρικούς». Φυσικά, στο μέλλον αυτό πρέπει να γίνει από ιατρικούς ειδικούς. Προτού ορίσουμε την νέα τελική ακρίβεια, ας σταθούμε στους ίδιους τους όρους στον παρακάτω πίνακα (τα χρώματα βασίζονται στο υπόμνημα της προηγούμενης πίτας):

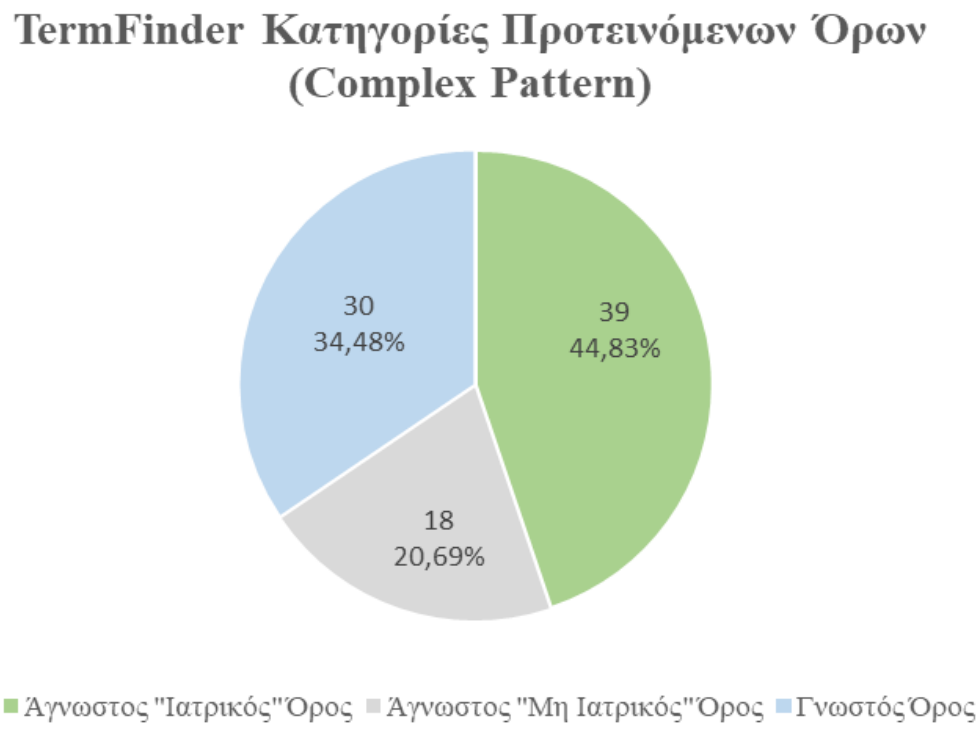
TermFinder **Προτεινόμενοι Όροι (Simple Pattern)**				
long - term autoimmune disease	*common auto - immune diseases*	*dry mouth*	*dry eyes*	*specific extant diagnostic test*
other organs systems	*other health problems*	*connective tissue disorder*	*other autoimmune diseases*	*systemic lupus erythematosus*
non-invasive diagnostic tests	*other autoimmune disorders*	*nervous system*	*primary symptoms*	*other symptoms*
dry skin	*vaginal dryness*	*chronic cough*	*joint pains*	*thyroid problems*
exact cause	*environmental trigger*	*primary sjögren*	*secondary sjögren*	*rheumatoid arthritis*
systemic sclerosis	*blood tests*	*specific antibodies*	*artificial tears*	*punctal plugs*
tear ducts	*saliva substitute*	*muscle pain*	*lip biopsy*	*henrik sjögren*
earlier descriptions	*primary form*	*secondary form*	*middle age*	*life expectancy*
sjögren	*syndrome*	*glands*	*disease*	*symptoms*
moisture	*dryness*	*pain*	*%*	*inflammation*
biopsy	*medications*	*sjs*	*ss*	*body*
lungs	*kidneys*	*fatigue*	*numbness*	*arms*
legs	*muscle*	*risk*	*lymphoma*	*combination*
genetics	*exposure*	*virus*	*bacterium*	*result*
ra	*sle*	*damages*	*diagnosis*	*lymphocytes*
treatment	*person*	*surgery*	*gum*	*sugar*
water	*ibuprofen*	*antihistamines*	*number*	*people*
population	*half*	*females*	*times*	*males*
anyone				

Πίνακας 4: Πίνακας με λίστα προτεινόμενων όρων από TermFinder σε φθίνουσα κατάταξη C-Value (Simple Pattern)

Κάτι ιδιαίτερα σημαντικό για τον Πίνακα με τους όρους είναι ότι οι όροι παρουσιάζονται με φθίνουσα σειρά κατάταξης ως προς το C-Value, δηλαδή οι πάνω όροι (και αν είναι στην ίδια γραμμη, οι πιο αριστερά) έχουν κριθεί ως πιθανότεροι όροι από τον TermFinder. Παρατηρούμε ότι όσο πιο ψηλά βρισκόμαστε τόσο πιο πολλούς «μπλέ» (δηλαδή γνωστούς) όρους βρίσκουμε. Άρα, αυτό δείχνει ότι ο αλγόριθμος μας δρα σωστά, αφού όσο ψηλότερο είναι το C-Value, τόσο πιθανότερο είναι να είναι πράγματι πραγματικός όρος.

Κάτι ακόμη θετικό, είναι ότι έχουμε πολλούς «πράσινους» όρους, όσο είμαστε ψηλά στον πίνακα. Οι πράσινοι όροι είναι άγνωστοι όροι που εντόπισε ο TermFinder, αλλά που μετά από δική μας ανάλυση αξιολογήσαμε ότι μάλλον αποτελούν ιατρικούς όρους. Το ότι οι «γκρί» όροι, δηλαδή οι άγνωστοι όροι που μας έδωσε ο TermFinder αλλά δε θεωρούμε ιατρικούς είναι λίγοι και συσσωρεύονται στο τέλος του πίνακα, μας δείχνει ότι στα πειράματα μας, θα πρέπει να λαμβάνουμε υπόψιν τους όρους με το υψηλότερο CValue, για να αποφύγουμε τους πολλούς «γκρι».

Συμπεριλαμβάνουμε την ίδια ανάλυση και για το πολύπλοκο συντακτικό μοτίβο για λόγους πληρότητας, αν και δεν αλλάζουν ιδιαίτερα οι παρατηρήσεις.



Σχήμα 11: Pie Chart Όρων που προτείνει ο TermFinder ως προς το είδος τους (Complex Pattern)

TermFinder **Προτεινόμενοι Όροι (Complex Pattern)**				
sjögren 's syndrome	*long - term autoimmune disease*	*blood tests for specific antibodies*	*common auto - immune diseases*	*dry mouth*
dry eyes	*primary sjögren 's syndrome*	*secondary sjögren 's syndrome*	*specific extant diagnostic test*	*number of earlier descriptions*
body 's moisture	*other organs systems*	*combination of genetics*	*other health problems*	*connective tissue disorder*
other autoimmune diseases	*systemic lupus erythematosus*	*biopsy of moisture*	*non-invasive diagnostic tests*	*person 's symptoms*
other autoimmune disorders	*nervous system*	*primary symptoms*	*other symptoms*	*dry skin*
vaginal dryness	*chronic cough*	*joint pains*	*thyroid problems*	*exact cause*
environmental trigger	*rheumatoid arthritis*	*systemic sclerosis*	*artificial tears*	*punctal plugs*
tear ducts	*saliva substitute*	*muscle pain*	*lip biopsy*	*henrik sjögren*
primary form	*secondary form*	*middle age*	*life expectancy*	*glands*
disease	*symptoms*	*dryness*	*pain*	*%*
inflammation	*biopsy*	*medications*	*sjs*	*ss*
lungs	*kidneys*	*fatigue*	*numbness*	*arms*
legs	*muscle*	*risk*	*lymphoma*	*exposure*
virus	*bacterium*	*result*	*ra*	*sle*
damages	*diagnosis*	*lymphocytes*	*treatment*	*surgery*
gum	*sugar*	*water*	*ibuprofen*	*antihistamines*
people	*population*	*half*	*females*	*times*
males	*anyone*			

Πίνακας 5: Πίνακας με λίστα προτεινόμενων όρων από TermFinder σε φθίνουσα κατάταξη C-Value (Complex Pattern)

Μετά από την παρουσίαση των ποσοστών ακρίβειας πριν ληφθούν υπόψιν οι ιατρικοί όροι που εντόπισε ο TermFinder (και κρίναμε εμείς ότι τελικά αποτελούν όρους), σειρά έχει να υπολογίσουμε τα ποσοστά ακρίβειας που προκύπτουν αν συμπεριληφθούν στους γνωστούς και οι ιατρικοί όροι και υπολογίσουμε την ακρίβεια με τον τύπο Τύπος 3: τελική ακρίβεια αντί για τον τύπο Τύπος 2: ακρίβεια

TermFinder (Τελική Ακρίβεια) Προτεινόμενοι Όροι	**Simple Pattern**	**Complex Pattern**
Γνωστοί Όροι που βρέθηκαν	29	30
Άγνωστοι Όροι που βρέθηκαν	62	57
Άγνωστοι Όροι που θεωρήθηκαν «Ιατρικοί»	41	69
Όλοι οι Όροι που βρέθηκαν (Γνωστοί + Άγνωστοι)	91	87
Τελική Ακρίβεια	76,92%	79,31%

Παρατηρούμε ότι αυξήσαμε τα ποσοστά μας από 31,87% σε 76,92% και από 34,48% σε 79,31%, όταν συμπεριλάβαμε στου ιατρικούς όρους και αυτού που προτείνει ο TermFinder και έχουμε αξιολογήσει εμείς εκ των υστέρων ότι είναι πιθανοί ιατρικοί όροι. Ανεξάρτητα από το γεγονός ότι τα ποσοστά αυτά είναι υποκειμενικά και θα θέλαμε στο μέλλον να αντικατασταθούν είτε από την ανάλυση ενός Ιατρικού ειδικού, είτε από τη σύνδεση με μία πραγματική ηλεκτρονική Ιατρική Βιβλιοθήκη, παρατηρούμε ότι ο TermFinder ανιχνεύει με παραπάνω από διπλάσια ακρίβεια όρους από ό,τι ένας άνθρωπος που κάνει μια βεβιασμένη ανάγνωση (και μάλιστα ο αλγόριθμος το πετυχαίνει σε γρηγορότερο χρονικό διάστημα).

5 Πειραματική Εφαρμογή Συστήματος

Έπειτα από την αξιολόγηση του συστήματος αναγνώρισης όρων, προβήκαμε σε πειράματα επί πραγματικών δεδομένων, ούτως ώστε να υπολογιστεί στην πράξη η αξία της *C-Value* καθώς και να αξιοποιηθεί για πραγματικούς σκοπούς η λειτουργία του *TermFinder*. Στο πλαίσιο αυτό, με δεδομένη την πρότερη ενασχόληση του εργαστηρίου με το project HarmonicSS[9] και την ύπαρξη έτοιμων πηγών δεδομένων, ασχοληθήκαμε με την εφαρμογή του *TermFinder* ως μέσου επέκτασης και υποβοήθησης της συνεχιζόμενης ιατρικής έρευνας που πραγματεύεται με το σύνδρομο Σιόγκρεν.

Στα πλαίσια της εργασίας αυτής, η βοηθητική αυτή λειτουργία έγινε με τη μορφή ξεχωριστών πειραμάτων στα οποία από τη μία εισάγαμε διαφορετικές εισόδους (π.χ. ιατρικά άρθρα, κλινικές δοκιμές, νοσοκομειακά δεδομένα) και από την άλλη τροποποιούσαμε τις παραμέτρους της εφαρμογής, σε όλα τα στάδια που αναλύσαμε στην προηγούμενη ενότητα. Προχωρούσαμε σε εκτέλεση της εφαρμογής πολλαπλές φορές για όλες τις παραλλαγές των προηγούμενων παραμέτρων και προωθούσαμε τις εξόδους στο Πρόγραμμα Δοκιμής και Αξιολόγησης Αποτελεσμάτων (Π.Δ.Α.Α) για να υπολογιστούν οι μετρικές αξιολόγησης. Στο τέλος, όμοια με τη μέθοδο που ακολουθήσαμε κατά την Αξιολόγηση του Συστήματος, αναλύαμε μόνοι μας οπτικά τους Άγνωστους Όρους που ήταν στην κορυφή ως προς το C-Value και μαζεύαμε έτσι νέους πιθανούς όρους για να εμπλουτίσουν μία ιατρική οντολογία.

Στην επόμενη ενότητα (5.1) περιγράφουμε σύντομα τα αρχεία εισόδων που αναφέραμε. Μετά από τις εκτελέσεις, προβήκαμε σε δική μας ανάλυση και σχολιασμό των αποτελεσμάτων και στην ενότητα που ακολουθεί (5.2) παρουσιάζονται τα σημαντικότερα ευρήματα, ενώ αναλυτικότερες λεπτομέρειες για όλα τα πειράματα ή καθένα από αυτά ξεχωριστά μπορούν να βρεθούν στην ενότητα (5.3). Παραπέμπουμε επίσης και στο παράρτημα (7) για τη λίστα με τους Προτεινόμενους Όρους.

5.1 Αρχεία Εισόδου – Datasets

Στις υποενότητες που ακολουθούν θα αναφερθούμε με συντομία στα δεδομένα εισόδου που τροφοδοτήσαμε στον TermFinder. Θα σταθούμε στα μεγέθη, τις ειδικές ανάγκες που είχαμε για καθένα από αυτά και τέλος, θα αναλύσουμε τις όποιες παραδοχές κάναμε επί αυτών.

[9] HarmonicSS Project, https://www.harmonicss.eu/

5.1.1 Μοντέλα Περιγραφής Δεδομένων Ασθενών - Dataset

Τα μοντέλα περιγραφής Δεδομένων Ασθενών είναι excel αρχεία τα οποία περιέχουν μεταδεδομένα που έχουν παραχθεί αυτόματα από το Cohort Metadata Extraction Tool από βάσεις και σχετικά έγγραφα του HarmonicSS. Περιέχουν καταγεγραμμένες παραμέτρους για τους ασθενείς ενός ιδρύματός καθώς και τις πιθανές τιμές των παραμέτρων αυτών.

Για να αναλύσουμε το έγγραφο, διακρίναμε 2 περιπτώσεις:

1) Η πρώτη ήταν να ασχοληθούμε μόνο με αυτά ακριβώς τα ζεύγη παραμέτρων τιμών. Να λάβουμε δηλαδή υπόψιν μόνο τα ονόματα των πεδίων (και τις τιμές τους). Δηλαδή να πάρουμε απευθείας τα προτεινόμενα ονόματα όρων όπως αναφέρονταν.

2) Η δεύτερη ήταν να επεκτείνουμε την αναζήτηση, συμπεριλαμβάνοντας εκτός από τη στήλη ονόματος πεδίων αλλά και την περιγραφή της και όμοια την περιγραφή και των τιμών. Με αυτόν τον τρόπο είχαμε ένα μεγαλύτερο σετ κειμένου που πιθανώς να έκρυβε κι άλλους όρους μέσα, πιθανά συγγενικούς: Για παράδειγμα, σε ένα excel είδαμε την τιμή «sicca», και στο description την περιγραφή «dry mouth and dry eyes». Ήδη εδώ ανιχνεύουμε δεξιά 2 πιθανούς όρους που χάναμε στην περίπτωση 1 («dry mouth», «dry eyes»). Θα αναφερόμαστε σε αυτή την περίπτωση στα πειράματα ως Δεδομένα+

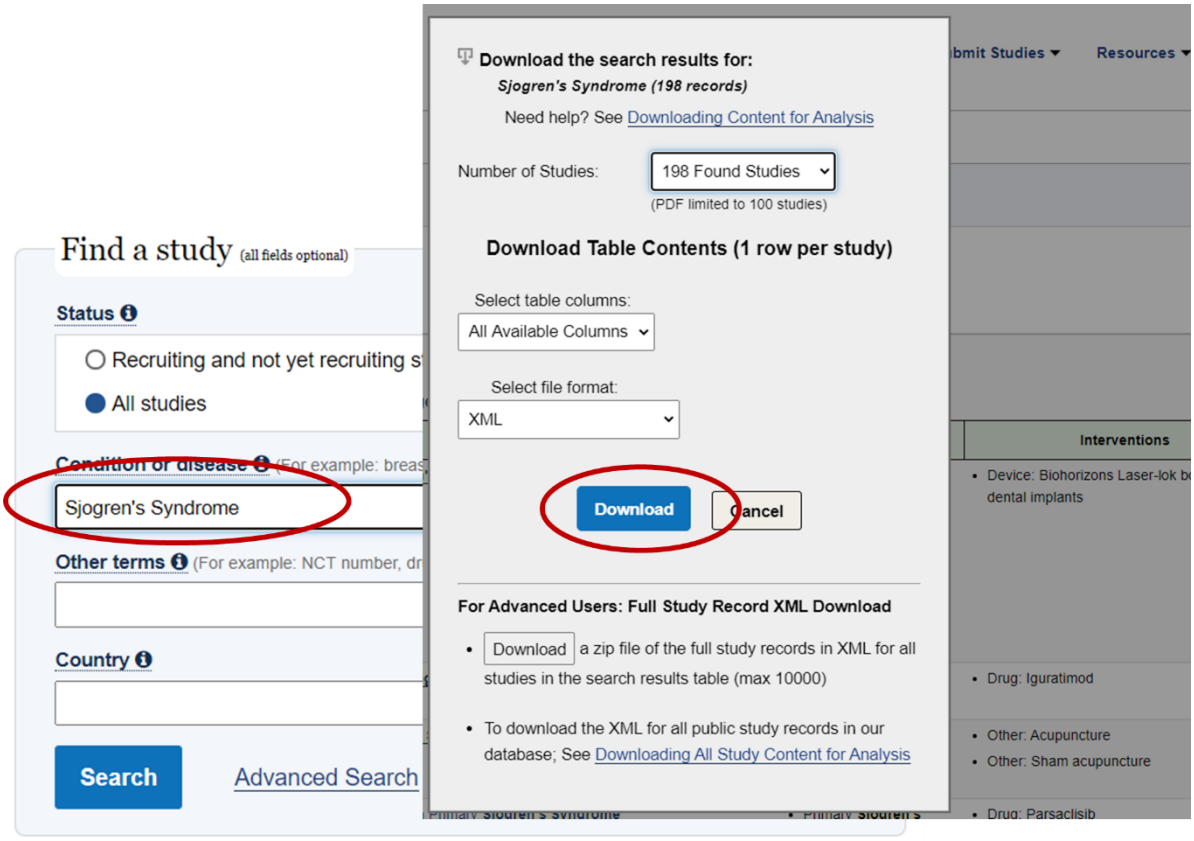
Το dataset μας περιείχε συνολικά 23 αρχεία excel, συνολικού μεγέθους 866KB. Προφανώς η πληροφορία περιορίστηκε σε πολύ μικρότερο υποσύνολο αυτού του μεγέθους, αφού λαμβάναμε το περιεχόμενο μόνο συγκεκριμένων στηλών σε καθεμία από τις 2 περιπτώσεις.

	A	B	C	D	E	F
1	ID	CATEGORY	SUB-CATEGORY	FIELD NAME	FIELD DESCRIPTION	DATA TYPE
2	A	Person	Person	PT	Patient study number -Person UID	
3	B	Person	Date of birth	GEBDAT	Date of birth-Date of birth	INTEGER
4	C	Person	Gender	GESLACHT	Gender -Gender	INTEGER
5	D	Pharmaceutical drugs	Glucocorticoids	PRED	Prednisone- Units of mesurement and conversion if it i	INTEGER
6	E	Pharmaceutical drugs	Glucocorticoids	MEFREQ01	Frequency-	INTEGER
7	F	Pharmaceutical drugs	Glucocorticoids	MEFREQEA01	Frequency units-	INTEGER
8	G	Pharmaceutical drugs	Glucocorticoids	MEFREQE01	Frequency per...-	INTEGER
9	H	IGNORE	IGNORE	MEFREQES01	Frequency other-	-
10	I	Pharmaceutical drugs	Glucocorticoids	MEDOSIS01	Medication dosage-	REAL NUMBER
11	J	Pharmaceutical drugs	Glucocorticoids	MEDOSISE01	Dosage units-	INTEGER
12	K	Pharmaceutical drugs	Glucocorticoids	MEDOSISES01	Dosage units other-	
13	L	Pharmaceutical drugs	Glucocorticoids	MESTARTD01	Medication startdate-	INTEGER
14	M	IGNORE	IGNORE	MESTOPD01	Medication stopdate-	-
15	N	Pharmaceutical drugs	Conventional DMARD	MTX	Methotrexate- Units of measurement and conversion if	INTEGER
16	O	Pharmaceutical drugs	Conventional DMARD	MEFREQ02	Frequency-	INTEGER
17	P	Pharmaceutical drugs	Conventional DMARD	MEFREQEA02	Frequency units-	INTEGER
18	Q	Pharmaceutical drugs	Conventional DMARD	MEFREQE02	Frequency per...-	INTEGER
19	R	IGNORE	IGNORE	MEFREQES02	Frequency other-	-
20	S	Pharmaceutical drugs	Conventional DMARD	MEDOSIS02	Medication dosage-	REAL NUMBER

Εικόνα 2: Παράδειγμα sheet XLS αρχείου για τα μοντέλα περιγραφής Δεδομένων Ασθενών

5.1.2 Περιγραφή Κλινικών Δοκιμών – Dataset

Η περιγραφή κλινικών δοκιμών προέρχεται από αναζήτηση πλήθους κλινικών δοκιμών (1 xml file ανά κλινική δοκιμή) από την ιστοσελίδα clinicaltrials.gov[10], φιλτράροντας ώστε να αφορά την ασθένεια του συνδρόμου Σιόγκρεν. Θεωρήσαμε δόκιμο να αναζητήσουμε και από εκεί όρους, καθώς πρόκειται για κείμενο με καθαρά ιατρικό σκοπό στο οποίο είναι αρκετά πιθανό να αναφέρονται ιατρικές ορολογίες. Επίσης, πρόκειται για ένα πιο δομημένο κείμενο από τα μοντέλα περιγραφής, αφού δεν είναι απλά κελιά σε ένα excel φίλε αλλά προέρχονται από δομημένο κείμενο.



Εικόνα 3: Η διαδικασία εξαγωγής των xml files για τα clinical trials.

Το dataset αυτό το εισάγουμε μετά από export όλων των xmls σε 1 μεγάλο txt αρχείο, συνολικού μεγέθους 783KB

5.1.3 Ιατρικά Άρθρα - Dataset

Τα ιατρικά άρθρα προέρχονταν από το σύνολο cord-19 το οποίο περιέχει συγκεντρωμένες όλες τις σχετιζόμενες με τον κορονοϊό έρευνες που είναι διαθέσιμες ελεύθερα online από το PubMed, τον World Health Organization και τα bioRxiv και medRxiv. Περισσότερες πληροφορίες για αυτά μπορούν να

[10] https://clinicaltrials.gov/

βρεθούν στην ίδια την ιστοσελίδα[11] τους . Θελήσαμε να δούμε πώς η εφαρμογή συσχετίζει το Sjogren Syndrome με τον νέο κορονοϊό, για αυτό και επιλέξαμε ένα υποσύνολο των άρθρων αυτών που περιείχε την λέξη "Sjogren". Για λόγους επιδόσεων της εφαρμογής αλλά και του εξαιρετικά μεγάλου μεγέθους των άρθρων, που συνεπαγόταν απαγορευτική διάρκεια των πειραμάτων (>24 ώρες), για τα δεδομένα αυτής της διπλωματικής επιλέξαμε ένα υποσύνολο και σε αυτά τα άρθρα, λαμβάνοντας 22 από αυτά, συνολικού μεγέθους 5.84 MB.

Εικόνα 4: Παράδειγμα JSON αρχείου από CORD-19 Dataset σχετικό με κάποιο medical paper

5.2 Μέθοδος Πειραμάτων – Τα Πειράματα

Όπως αναφέρθηκε και νωρίτερα, διενεργήθηκαν συνολικά πειράματα σε 3 διαφορετικές εισόδους. Η μέθοδος που ακολουθήθηκε ήταν κοινή και περιλαμβάνει τα βήματα της φόρτωσης του *Αρχείου Διαμόρφωσης* (.yml) και επανάληψης των εκτελέσεων για κάθε συνδυασμό διαφορετικών παραμέτρων στο pipeline του *TermFinder*.

Το Αρχείο Διαμόρφωσης, όπως αναφέραμε και στην ενότητα (3.3.1) σχεδιάστηκε ώστε να αποτελεί την εργαλειοθήκη μας για τις παραμετροποιήσεις με τη βοήθεια του Προγράμματος Διαχείρισης Υποσυστημάτων, για αυτό αποτέλεσε και το κύριο εργαλείο ελέγχου τον πειραμάτων μας. Στο πλαίσιο της διπλωματικής εργασίας αυτής, θέσαμε ως παραμέτρους και δοκιμάσαμε διαφορετικές τιμές για τις ακόλουθες μεταβλητές που επεμβαίνουν στην εκτέλεση του προγράμματος:

[11] Cord-19 Related Data: https://www.semanticscholar.org/cord19/download

1) Το μοτίβο Part-Of-Speech αναζήτησης όρων (Pattern)

2) Το Αρχείο Αναζητούμενων Πραγματικών όρων (Evaluator)

Άλλες παραλλαγές οι οποίες θα μπορούσαν να είχαν μελετηθεί θα μπορούσαν να αφορούν τη παραμετροποίηση του stopword list (από πιο χαλαρό σε πιο αυστηρό) ή του edit-distance elasticity για ομαδοποίηση όρων (το οποίο έχει τεθεί στον 1 διαφορετικό χαρακτήρα για αποφυγή false positives). Οι προηγούμενες διαφοροποιήσεις είναι διαθέσιμες με αλλαγές στις παραμέτρους του configuration (stoplist_path, edit_distance). Άλλες ενδιαφέρουσες παράμετροι θα μπορούσαν να προκαλούν διαφοροποίηση των εσωτερικών εργαλείων tokenization και POS-Tagging ώστε να συγκριθούν οι επιδόσεις τους.

Στο τέλος του pipeline του *TermFinder*, τα αποτελέσματα τροφοδοτούνται στον *Αλγόριθμο Δοκιμής και Αξιολόγησης Αποτελεσμάτων* ο οποίος παράγει σε μορφή txt τις μετρικές που θα παρουσιάσουμε στην επόμενη ενότητα.

Συνολικά, εκτελέσαμε πειράματα για 2 παραλλαγές μοτίβων POS-Tagging (Patterns) τα οποία είναι τα εξής:

1. (((NN|JJ))*NN)
2. (((((NN|JJ))*NN) IN (((NN|JJ))*NN))|
   ((NN|JJ)*NN POS (NN|JJ)*NN))|
   (((NN|JJ))*NN)

Η 1η θα αναφέρεται στο εξής ως Simple Pattern ενώ η 2η ως Complex Pattern. Παραπέμπουμε στην ενότητα 62 για αναλυτικότερη επεξήγηση τους.

Επιπλέον, διαφοροποιήσαμε τα πειράματα μας και ως προς το Αρχείο Αναζητούμενων Πραγματικών όρων. Ενώ η πηγή παρέμεινε σταθερή, δηλαδή επρόκειτο για την οντολογία του HarmonicSS, οι αναζητούμενοι όροι διαφοροποιήθηκαν σε 4 διαφορετικές σημασιολογικά οντότητες:

0) Ιατρικοί όροι - Πάσης φύσεως (Αναζήτηση Κάθε Όρου)

1) Ιατρικοί όροι – Ασθένειες (Diseases)

2) Ιατρικοί όροι – Φάρμακα (Drugs)

3) Ιατρικοί όροι – Δημογραφικά Δεδομένα (Demographics)

Η 1η είναι ένα υπερσύνολο των άλλων 3, ενώ η τελευταία επρόκειτο για την πιο μικρή και ίσως πιο γενικής φύσεως κατηγορία όρων. Τα Diseases και τα Drugs πιστεύουμε ότι αποτελούν πιο σημαντικές

και συνηθισμένες περιπτώσεις όρων που θα χρειαζόταν ένας γιατρός ή ένας ερευνητής. Για χάριν συντομίας στο παρόν κείμενο, στα επόμενα κεφάλαια θα υπάρξουν τα αποτελέσματα για την περίπτωση 0, ενώ για τις άλλες 3 θα προβούμε σε ανάλυση και κατηγοριοποίηση επί τον αποτελεσμάτων του 0.

Από τα σύνολα των όρων αυτών που προέρχονται από το HarmonicSS, στα πειράματα μας χρησιμοποιούμε ως μέτρο αξιολόγησης και «χρυσό πρότυπο» τους επονομαζόμενους Υπαρκτούς Γνωστούς Όρους, δηλαδή όσους εκ των όρων που αναφέραμε παραπάνω υπάρχουν έστω και 1 φορά στα κείμενα εισόδου. Δηλαδή δεν έχει νόημα, για παράδειγμα στην περίπτωση 3, για τα φάρμακα να αναζητούμε το C-Value termhood για την «*Sulphasalazine»* αν δεν υπάρχει πουθενά αναφορά της π.χ. στο Dataset των Δεδομένων Ασθενών με Σύνδρομο Σιόγκρεν. Τέλος, συνειδητοποιήσαμε ότι θα μπορούσαμε να καταστήσουμε πλήρως αξιόπιστη την αξιολόγηση όχι μόνο στην περίπτωση που αυτό που ψάχνουμε υπάρχει ακριβώς, αλλά και στην περίπτωση που συναντάται με κάποια αντίστοιχη μορφή, άρα οι *Υπαρκτοί Γνωστοί Όροι*, λαμβάνουν και αυτό υπόψιν κατά την αξιολόγηση.

Άρα, συνολικά, όπως φαίνεται και παραπάνω, για κάθε περίπτωση εισόδου, που όπως είπαμε και πριν είναι τα δεδομένα ασθενών, οι κλινικές δοκιμές και τα επιστημονικά άρθρα, εκτελέσαμε 2 Patterns * 6 Evaluators = 12 διαφορετικές εκτελέσεις του αλγορίθμου εκτελώντας κάθε συνδυασμό Μοτίβου με Αρχείο Αναζητούμενων Πραγματικών όρων. Στη συνέχεια θα δούμε μία εποπτική αξιολόγηση των αποτελεσμάτων ενώ θα έχουμε 1 ενότητα για κάθε περίπτωση εισόδου, στην οποία θα βλέπουμε τις 12 εκτελέσεις που αντιστοιχούν σε αυτή.

5.3 Σύγκριση Αποτελεσμάτων

Όμοια με τη δοκιμαστική εκτέλεση στο κείμενο της Wikipedia, και η διαδικασία των πειραμάτων έδωσε τη δυνατότητα να βρεθούν νέοι όροι, μέσα από ανάλυση των κορυφαίων αγνώστων όρων ως προς το C-Value τους. Εδώ, χρησιμοποιούμε ως αρχείο Γνωστών Όρων την οντολογία του HarmonicSS [4], το οποίο όμως σημαίνει ότι μπορεί σε κάποια κείμενα να μην υπάρχει ουδεμία αναφορά σε κάποιους εκ των όρων. Στις πρώτες μας μετρήσεις, χωρίς τη μέθοδο αναζήτησης των Υπαρκτών Γνωστών Όρων, τα ποσοστά επιτυχίας του αλγορίθμου *TermFinder* ήταν εξαιρετικά χαμηλά, χωρίς να ευθύνεται ο αλγόριθμος, επομένως τις παραλείψαμε ως αβάσιμες. Τα πειράματα που ακολουθούν συμβαδίζουν με την τελευταία έκδοση του αλγορίθμου τη στιγμή της συγγραφής της παρούσας εργασίας.

Η εικόνα που αποκομίσαμε από τα πειράματα μας ήταν ότι η ακρίβεια του εντοπισμού των όρων ποικίλει ανάλογα με το επιλεγμένο μοτίβο λέξεων αλλά και με το μέγεθος της εισόδου που βάζουμε

στον αλγόριθμο. Ενδιαφέρουσα είναι η παρατήρηση ότι το Simple Pattern είχε συνολικά καλύτερες επιδόσεις από το Complex Pattern κατά %, γεγονός που ευθύνεται πιθανώς από το γεγονός ότι στους evaluators υπήρχαν λιγότεροι όροι με σύνθετη μορφή από απλή και άρα οι περισσότεροι σύνθετοι όροι του TermFinder απορρίπτονταν.

Παραθέτουμε στη συνέχεια τους πίνακες με τις επιδόσεις των Simple και Complex Pattern για την περίπτωση Evaluator (1) δηλαδή για Ιατρικούς Όρους Πάσης φύσεως (Αναζήτηση Κάθε Όρου) για τα 4 πειράματα που εκτελέσαμε. Θα αναφερόμαστε σε αυτά ως εξής:

1) Μοντέλα Περιγραφής Δεδομένων Ασθενών: Αναζήτηση στα excel με τα Μοντέλα Περιγραφής Δεδομένων Ασθενών στις στήλες Field Name και Value
2) Μοντέλα Περιγραφής Δεδομένων Ασθενών: Αναζήτηση στα excel με τα Μοντέλα Περιγραφής Δεδομένων Ασθενών στις στήλες Field Name, Field Description, Value και Value Description
3) Περιγραφή Κλινικών Δοκιμών: Αναζήτηση στο txt με το export από την περιγραφή όλων των μελετών για τις κλινικές δοκιμές.
4) Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren: Αναζήτηση στα json ιατρικά papers και μελέτες επί του COVID-19 με φιλτράρισμα ώστε να αναφέρονται στο Sjogren Syndrome και περισυλλογή μόνο δεδομένων από τα «body-text» values.

Πρώτα θα ξεκινήσουμε με την εύρεση υποσύνολο των όρων του λεξιλογίου που παρείχαμε ως evaluator που πράγματι υπάρχει στα αρχικά δεδομένα εισόδου για τα 4 πειράματα. Αρχικά, στο export των όρων οντολογίας του HarmonicSS ξεκινάμε με μία βάση από 653 όρους. Στη συνέχεια θα βρούμε σε μία πρώτη φάση πόσοι από αυτούς τους 653 όρους υπάρχουν έστω και 1 φορά στα πειραματικά κείμενα εισόδου.

Παρατηρούμε ότι τα ποσοστά Γνωστών Όρων που υπάρχουν στα κείμενα μελέτης είναι αρκετά μικρότερα από τους αρχικούς γνωστούς όρους.

Τα Μοντέλα Περιγραφής Δεδομένων Ασθενών, φαίνεται να κρύβουν το μεγαλύτερο αριθμό όρων, ειδικά αν λάβουμε υπόψη ότι έχουν το μικρότερο και πιο συμπυκνωμένο dataset. Κοντά στους 200 μπορούν να βρεθούν και στη Περιγραφή Κλινικών Δοκιμών, μόνο που εδώ σε σύγκριση με τα μοντέλα περιγραφής ο αλγόριθμος είχε μεγαλύτερο κείμενο να επεξεργαστεί. Τα κείμενα του Covid που κάνουν αναφορά στο Sjogren, περιέχουν αρκετά μικρό αριθμό όρων σχετικών με το Sjogren αν λάβουμε υπόψιν το αρκετά μεγάλο μέγεθος τους, αν και είναι αναμενόμενο αφού δεν εξειδικεύονται στο Sjogren.

Θα εκτελούμε το πείραμα επί των top 10.000 υποψήφιων όρων, αλλά στη συνέχεια θα επιμείνουμε κυρίως στις κατατάξεις των Υπαρκτών Όρων εντός των προηγουμένων, για να υπολογίσουμε την ισχύ του αλγορίθμου μας. Θα σταθούμε και στους υπόλοιπους όρους που ανιχνεύουμε για να δούμε σε ένα πρώτο στάδιο αν είναι πράγματι όροι, εστιάζοντας σε αυτούς με το μεγαλύτερο C-Value.

Ακολουθούν πίνακες με την πρωτογενή ανάλυση ως προς το αν ανιχνεύτηκαν τελικά και πόσοι Γνωστοί Όροι στους top 10.000 υποψήφιους όρους που προτείνει ο TermFinder, για τις 2 περιπτώσεις μοτίβων POS-Sequences που αναφέραμε στο (5.2), την απλή (Simple) και τη σύνθετη (Complex).

TermFinder (Simple Pattern) Top 10.000 Προτεινόμενοι Όροι	**Μοντέλα Περιγραφής Δεδομένων Ασθενών**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+**	**Περιγραφή Κλινικών Δοκιμών**	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**
Γνωστοί Όροι που Υπάρχουν	147	282	193	203
Γνωστοί Όροι στο Top 10.000	147	278	180	134
Ποσοστό Επιτυχίας	100.00%	98.58%	93.26%	66.01 %

Πίνακας 6: Πρωτογενής αξιολόγηση προτεινόμενων όρων αλγορίθμου με απλό μοτίβο POS Tagging

TermFinder (Complex Pattern) Top 10.000 Προτεινόμενοι Όροι	**Μοντέλα Περιγραφής Δεδομένων Ασθενών**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+**	**Περιγραφή Κλινικών Δοκιμών**	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**
Γνωστοί Όροι που Υπάρχουν	143	285	187	188
Γνωστοί Όροι στο Top 10.000	143	281	148	121
Ποσοστό Επιτυχίας	100.00%	98.60%	79.14%	64.36%

Πίνακας 7: Πρωτογενής αξιολόγηση προτεινόμενων όρων αλγορίθμου με σύνθετο μοτίβο POS Tagging

5.3.1 Μοντέλα Περιγραφής Δεδομένων Ασθενών – Ανάλυση

Θα αναλύσουμε τώρα πιο διεξοδικά τα αποτελέσματα ανά περίπτωση input. Στην 1η στήλη (χωρίς το + σύμβολο) που σχετίζεται με parsing μόνο των στηλών field name και value, ενώ η 2η (με το + σύμβολο) που σχετίζεται με parsing των στηλών field name, value, description, value description. Η 3η και η 4η στήλη είναι επαναλήψεις των στηλών 1 και 2, απλά για εκτέλεση του TermFinder με το Complex pattern, για αυτό και έχει τεθεί ένα C στο τέλος των τίτλων τους.

Πρώτα, θα σταθούμε με μια γενική εποπτική εικόνα της μέσης κατάταξης των επικυρωμένων όρων ως προς το Top% Μέσης Κατάταξης:

$$Top\%\ Μέσης\ Κατάταξης = \frac{rank\_όρου(avg\_CValue(set\ of\ Known\ Terms))}{count(set\ of\ All\ Found\ Terms)}$$

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Μοντέλα Περιγραφής Δεδομένων Ασθενών**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών C**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+ C**
Μέση Κατάταξη Γνωστών Όρων	578	607	649	608
Συνολικοί Προτεινόμενοι Όροι αν <10000	1945	3247	2011	3430
Top% Μέσης Κατάταξης Γνωστών Όρων	29.7%	18.7%	32.3%	17.7%

Πίνακας 8: Κατατάξεις ως προς C-Value Γνωστών Όρων στο σύνολο των Όρων του TermFinder – Πείραμα 1

Ο λόγος που ασχολούμαστε με την άνωθι μετρική είναι ότι οι Συνολικοί Προτεινόμενοι Όροι είναι στις τάξεις των χιλιάδων οπότε είναι δύσκολο να βγάλουμε γενικά συμπεράσματα για τη σχετική τους θέση με απλή αναπαράσταση τους σε tables με τους όρους (όπως κάναμε στην ενότητα 4).

Παρατηρούμε ότι την καλύτερη κατάταξη τη λαμβάνουμε στο Μοντέλο Περιγραφής Δεδομένων Ασθενών+ με σύνθετο συντακτικό μοτίβο, άρα θα ασχοληθούμε με αυτόν όταν δούμε τους κορυφαίους άγνωστους που βρίσκει ο TermFinder, λίγο πιο κάτω.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Μοντέλα Περιγραφής Δεδομένων Ασθενών**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών C**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών+ C**
Μέγιστο CValue Γνωστών Όρων	7.3366	67.418	7.3366	67.418
Μέσο CValue Γνωστών Όρων	1.1509	4.8209	1.1837	4.7295
Ελάχιστο CValue Γνωστών Όρων	0.10000	0.10000	0.10000	0.10000

Πίνακας 9: Μέγιστο-Μέσο-Ελάχιστο CValue για Γνωστούς Όρους – Πείραμα 1

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Μοντέλα Περιγραφής Δεδομένων Ασθενών**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών +**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών C**	**Μοντέλα Περιγραφής Δεδομένων Ασθενών + C**
Μέγιστο CValue Γενικά	11.897	430.00	11.986	430.13
Μέσο CValue Γενικά	0.81666	2.2317	0.89716	2.4761
Ελάχιστο CValue Γενικά	0.10000	0.10000	0.10000	0.10000

Πίνακας 10: Μέγιστο-Μέσο-Ελάχιστο CValue για όλους τους Όρους Μαζί – Πείραμα 1

Όσον αφορά τους 2 προηγούμενους πίνακες, ο λόγος που τους παραθέτουμε εδώ είναι για να συγκρίνουμε τις βασικές μετρικές ελαχίστου μέσου και μέγιστου CValue, ανάμεσα στους Γνωστούς Όρους και όλους τους Προτεινόμενους Όρους. Συγκεκριμένα, φαίνεται ότι στα Μοντέλα Περιγραφής Δεδομένων Ασθενών+ Complex Pattern έχουμε σχεδόν διπλό μέσο CValue στους Γνωστούς όρους.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Περιγραφή Δεδομένων Ασθενών**	**Περιγραφή Δεδομένων Ασθενών +**	**Περιγραφή Δεδομένων Ασθενών C**	**Περιγραφή Δεδομένων Ασθενών + C**
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέγιστο CValue (Πλήθος)	15	3	18	7
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέσο CValue (Πλήθος)	376	202	508	252
Άγνωστοι όροι άνω από Γνωστό Όρο με Ελάχιστο CValue (Πλήθος)	1786	2926	1855	3097

Πίνακας 11: Πλήθος Άγνωστων όρων με σχετική κατάταξη σε σχέση ως προς τους Γνωστούς – Πείραμα 1

Ο παραπάνω πίνακας χρησιμεύει ώστε να βρούμε άγνωστους όρους άνω από τον καλύτερο γνωστό όρο μας ως προς το CValue. Περιορίζουμε τη μελέτη μας μόνο στους όρους άνω από Γνωστό Όρο με Μέσο CValue που έχουν επαρκές αλλά όχι υπερβολικό πλήθος αφού αυτοί είναι πιθανοί όροι. Η λίστα με τους top 100 όρους είναι στο παράρτημα, ενώ εδώ βλέπουμε 5 ενδεικτικούς «Άγνωστους» όρους ανά κατηγορία που δεν υπήρχαν στην οντολογία και προτάθηκαν από τον TermFinder:

Diseases	**Drugs/Therapies**	**Lab/Tests**	**Other**
hospital anxiety	*serum complement*	*schirmer 's test*	*reference model term*
urogenital system disease	*antirheumatic drug*	*dosage units*	*unit of measurement*
anti-hepatitis c virus	*antibodies*	*ocular staining score*	*quality of life assessment*
lymphatic system disease	*dosages for azathioprine*	*salivary gland biopsy*	*rose bengal staining*
renal disease	*chemotherapy*	*medication dosage*	*depression scale*

Πίνακας 12: Ενδεικτικοί προτεινόμενοι όροι που δεν υπάρχουν στο HarmonicSS ontology – Πείραμα 1

5.3.2 Περιγραφή Κλινικών Δοκιμών – Ανάλυση

Στην περίπτωση των κλινικών δοκιμών, έχουμε 2 περιπτώσεις, την 1η με το μοτίβο Simple και τη δεύτερη με το Complex. Πρώτα, θα σταθούμε με μια γενική εποπτική εικόνα της μέσης κατάταξης των επικυρωμένων όρων ως προς το Top% Μέσης Κατάταξης:

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Περιγραφή Κλινικών Δοκιμών**	**Περιγραφή Κλινικών Δοκιμών C**
Μέση Κατάταξη Γνωστών Όρων	1314	1425
Συνολικοί Προτεινόμενοι Όροι αν <10000	10000 (πραγμ. 11297)	10000 (πραγμ. 13052)
Top% Μέσης Κατάταξης Γνωστών Όρων	13.1%	14.25%

Πίνακας 13: Κατατάξεις ως προς C-Value Γνωστών Όρων στο σύνολο των Όρων του TermFinder – Κλινικές Δοκιμές – Πείραμα 2

Παρατηρούμε ότι έχουμε υψηλότερη κατάταξη με σύνθετο συντακτικό μοτίβο, άρα θα ασχοληθούμε με αυτό όταν δούμε τους κορυφαίους άγνωστους που βρίσκει ο TermFinder, λίγο πιο κάτω.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Περιγραφή Κλινικών Δοκιμών**	**Περιγραφή Κλινικών Δοκιμών C**
Μέγιστο CValue Γνωστών Όρων	187.05	187.53
Μέσο CValue Γνωστών Όρων	5.8874	6.5350
Ελάχιστο CValue Γνωστών Όρων	0.20000	0.79314

Πίνακας 14: Μέγιστο-Μέσο-Ελάχιστο CValue για Γνωστούς Όρους – Πείραμα 2

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Περιγραφή Κλινικών Δοκιμών**	**Περιγραφή Κλινικών Δοκιμών C**
Μέγιστο CValue Γενικά	215.22	564.55
Μέσο CValue Γενικά	2.2181	2.9870
Ελάχιστο CValue Γενικά	0.10000	0.79314

Πίνακας 15: Μέγιστο-Μέσο-Ελάχιστο CValue για Γνωστούς και Άγνωστους Όρους Μαζί – Πείραμα 2

Όσον αφορά τους 2 προηγούμενους πίνακες, ο λόγος που τους παραθέτουμε εδώ είναι για να συγκρίνουμε τις βασικές μετρικές ελαχίστου μέσου και μέγιστου CValue, ανάμεσα στους Γνωστούς Όρους και όλους τους Προτεινόμενους Όρους. Συγκεκριμένα, φαίνεται ότι στο σύνθετο συντακτικό μοτίβο έχουμε παραπάνω από διπλό μέσο CValue στους Γνωστούς όρους.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Περιγραφή Κλινικών Δοκιμών**	**Περιγραφή Κλινικών Δοκιμών C**
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέγιστο CValue (Πλήθος)	4	10
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέσο CValue (Πλήθος)	492	514
Άγνωστοι όροι άνω από Γνωστό Όρο με Ελάχιστο CValue (Πλήθος)	9685	9765

Πίνακας 16: Πλήθος Άγνωστων όρων με σχετική κατάταξη σε σχέση ως προς τους Γνωστούς – Πείραμα 2

Ο παραπάνω πίνακας χρησιμεύει ώστε να βρούμε άγνωστους όρους άνω από τον καλύτερο γνωστό όρο μας ως προς το CValue. Η λίστα με τους top 100 όρους υπάρχει στο παράρτημα, ενώ εδώ βλέπουμε 5 ενδεικτικούς «Άγνωστους» όρους ανά κατηγορία που δεν υπήρχαν στην οντολογία και προτάθηκαν από τον TermFinder:

Diseases	Drugs/Therapies	Lab/Tests	Other
hepatitis b	*oral health*	*inclusion criteria*	*informed consent*
lung diseases	*mycophenolate mofetil*	*exclusion criteria*	*quality of life*
hepatitis c	*placebo*	*clinical trial*	*european consensus criteria*
active infection	*dose level of sar441344*	*classification criteria*	*pilot study*
ocular disease	*artificial tears*	*side effects*	*healthy volunteers*

Πίνακας 17: Ενδεικτικοί προτεινόμενοι όροι που δεν υπάρχουν στο HarmonicSS ontology – Πείραμα 2

5.3.3 Ιατρικά Άρθρα - Ανάλυση

Για τα επιστημονικά ιατρικά άρθρα, σχετιζόμενα με COVID + Sjogren Syndrome, έχουμε την 1η περίπτωση με μοτίβο Simple και τη δεύτερη με Complex. Πρώτα, θα σταθούμε με μια γενική εποπτική εικόνα της μέσης κατάταξης των επικυρωμένων όρων ως προς το Top% Μέσης Κατάταξης:

TermFinder Top 10.000 Προτεινόμενοι Όροι	Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren	Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren
Μέση Κατάταξη Γνωστών Όρων	1397	1425
Συνολικοί Προτεινόμενοι Όροι αν <10000	10000 (πραγμ. 28922)	10000 (πραγμ. 34159)
Top% Μέσης Κατάταξης Γνωστών Όρων	14.0%	14.3%

Πίνακας 18: Κατατάξεις ως προς C-Value Γνωστών Όρων στο σύνολο των Όρων του TermFinder – Κλινικές Δοκιμές – Πείραμα 3

Παρατηρούμε ότι έχουμε υψηλότερη κατάταξη με σύνθετο συντακτικό μοτίβο, άρα θα ασχοληθούμε με αυτό όταν δούμε τους κορυφαίους άγνωστους που βρίσκει ο TermFinder, λίγο πιο κάτω.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**
Μέγιστο CValue Γνωστών Όρων	431.71	432.93
Μέσο CValue Γνωστών Όρων	8.5059	9.9593
Ελάχιστο CValue Γνωστών Όρων	1.1986	1.5863

Πίνακας 19: Μέγιστο-Μέσο-Ελάχιστο CValue για Γνωστούς Όρους – Πείραμα 3

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**
Μέγιστο CValue Γενικά	431.71	432.93
Μέσο CValue Γενικά	3.7354	4.0923
Ελάχιστο CValue Γενικά	1.1986	1.5863

Πίνακας 20: Μέγιστο-Μέσο-Ελάχιστο CValue για Γνωστούς και Άγνωστους Όρους Μαζί – Πείραμα 3

Όσον αφορά τους 2 προηγούμενους πίνακες, ο λόγος που τους παραθέτουμε εδώ είναι για να συγκρίνουμε τις βασικές μετρικές ελαχίστου μέσου και μέγιστου CValue, ανάμεσα στους Γνωστούς Όρους και όλους τους Προτεινόμενους Όρους. Συγκεκριμένα, φαίνεται ότι στο σύνθετο συντακτικό μοτίβο έχουμε παραπάνω από διπλό μέσο CValue στους Γνωστούς όρους.

TermFinder Top 10.000 Προτεινόμενοι Όροι	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**	**Ιατρικά Άρθρα (Cord-19) σχετικά με Sjogren**
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέγιστο CValue (Πλήθος)	0	0
Άγνωστοι όροι άνω από Γνωστό Όρο με Μέσο CValue (Πλήθος)	686	599
Άγνωστοι όροι άνω από Γνωστό Όρο με Ελάχιστο CValue (Πλήθος)	9802	9842

Πίνακας 21: Πλήθος Άγνωστων όρων με σχετική κατάταξη σε σχέση ως προς τους Γνωστούς – Πείραμα 3

Ο παραπάνω πίνακας χρησιμεύει ώστε να βρούμε άγνωστους όρους άνω από τον καλύτερο γνωστό όρο μας ως προς το CValue. Όπως στην ενότητα (63) αναλύαμε όλους τους όρους και βρίσκαμε στις ψηλότερες θέσεις άγνωστους όρους που ήταν ιατρικοί, έτσι κι εδώ, επειδή έχουμε πολύ μεγάλο αριθμό από όρους που βρίσκουμε, θα περιορίσουμε τη μελέτη μας μόνο στους όρους άνω από Γνωστό Όρο με Μέσο CValue που έχουν επαρκές αλλά όχι υπερβολικό πλήθος. Η λίστα με τους top 100 όρους υπάρχει στο παράρτημα, ενώ εδώ βλέπουμε 5 ενδεικτικούς «Άγνωστους» όρους ανά κατηγορία που δεν υπήρχαν στην οντολογία και προτάθηκαν από τον TermFinder:

Diseases	**Drugs/Therapies**	**Lab/Tests**	**Other**
acute myeloid leukemia	*standard ifn*	*marrow biopsy*	*b cell*
iron deficiency	*combination therapy*	*% of cases*	*age group*
pneumonia	*antibiotic therapy*	*serum ferritin*	*viral load*
virus	*induction chemotherapy*	*lung biopsy*	*blood donors*
hodgkin 's lymphoma	*antiviral drugs*	*chest radiographs*	*clinical response*

Πίνακας 22: Ενδεικτικοί προτεινόμενοι όροι που δεν υπάρχουν στο HarmonicSS ontology – Πείραμα 3

6 Σύνοψη

Στα πλαίσια αυτής της εργασίας, μελετήσαμε σε πρώτη φάση τη βιβλιογραφία που υπάρχει γύρω από την Αυτόματη Εξαγωγή Όρων. Με βάση το σχετικό θεωρητικό υπόβαθρο, δημιουργήσαμε μία εφαρμογή, ονομαζόμενη *TermFinder*, η οποία υλοποιεί τον αλγόριθμο στατιστικής εύρεσης όρων C-Value.

Προκειμένου να αξιολογήσουμε την αποτελεσματικότητα της, προβήκαμε σε μια πρώτη δοκιμαστική εκτέλεση της εφαρμογής έχοντας ως αρχείο εισόδου τις πρώτες 5 παραγράφους της εγγραφής της Wikipedia για το σύνδρομο Sjogren. Μιας και αποτελούσε ένα μικρό και ελεγχόμενο δείγμα, εντοπίσαμε με απλή ανάγνωση έναν μικρό αριθμό από ιατρικούς όρους (31), τους οποίους εξετάσαμε αν η εφαρμογή μας έβρισκε αυτόματα. Διαπιστώσαμε ότι ο *TermFinder* πετύχαινε ένα ποσοστό ανάκλησης 93.75% των προεντοπισμένων όρων αλλά ακρίβεια 34,48%, αν συμπεριλάβουμε όλους τους άγνωστους όρους που εντόπισε η εφαρμογή. Η τελική ακρίβεια βελτιώθηκε σε 79,31% όταν με δεύτερη αξιολόγηση αναγνωρίσαμε ένα υποσύνολο των άγνωστων όρων ως ιατρικούς.

Με δεδομένη την αποδεκτή αυτή ακρίβεια και ανάκληση της εφαρμογής μας, εξετάσαμε πειραματικά την αξιοποίηση της ως πρόγραμμα υποβοήθησης στην εύρεση ιατρικών όρων σχετιζόμενων με το σύνδρομο Sjogren. Η εφαρμογή τροφοδοτήθηκε με 3 διαφορετικά είδη κειμένων σχετικά με το σύνδρομο, που αφορούσαν μελέτες, νοσοκομειακές σημειώσεις και κλινικές οδηγίες και υπολόγισε τους πιθανότερους ιατρικούς όρους στα κείμενα. Τα αποτελέσματα συλλέχτηκαν και επί αυτών αναζητήσαμε ήδη γνωστούς όρους προερχόμενους από την οντολογία HarmonicSS, ώστε να εκτιμήσουμε πόσους έβρισκε στα κείμενα και πόσο ψηλά τους κατέτασσε ως προς το C-Value τους. Η αξία της εφαρμογής φάνηκε όμως από τις προτάσεις που μας έδωσε για πιθανούς όρους που δεν είχαμε στην οντολογία, άγνωστοι όροι δηλαδή που είχαν καταταχτεί ως προς το C-Value τους ψηλότερα από ήδη γνωστούς και άρα ήταν πιθανό να είναι ιατρικοί.

Κατά την μελέτη των αποτελεσμάτων μας, επιλέξαμε ενδεικτικά από αυτά ιατρικούς όρους που αναφέρονται σε ασθένειες, θεραπείες και κλινικές δοκιμές και τους παραθέσαμε για μελέτη από τον αναγνώστη. Τα αποτελέσματα φαίνονται ενθαρρυντικά αφού αρκετές προτάσεις είναι πράγματι ιατρικού περιεχομένου και υποστηρίζουν την αρχική υπόθεση της εργασίας ότι μπορούν να χρησιμοποιηθούν στατιστικές μέθοδοι όπως η C-Value για αυτόματο εντοπισμό ιατρικών όρων. Είναι εμφανές ότι η αξιολόγηση από μεριάς μας για τις κατηγορίες των ευρεθέντων άγνωστων όρων δεν

μπορεί να θεωρηθεί επαρκής λόγω μη ειδίκευσης μας στον ιατρικό τομέα, επομένως μία αξιολόγηση των αποτελεσμάτων από κάποιον ιατρικό ερευνητή θα ήταν επιθυμητή.

Πιθανή πορεία για μελλοντική βελτίωση της εφαρμογής είναι μία δεύτερη αξιολόγηση των άγνωστων όρων μέσα από κάποιου είδους αυτοματοποιημένη αναζήτηση τους σε online ιατρικές βιβλιοθήκες (PubMed, GENIA). Μια άλλη ενδιαφέρουσα δίοδος που δεν είχαμε την ευκαιρία να ακολουθήσουμε στα πλαίσια της διπλωματικής, θα σχετιζόταν πιθανώς με την αξιοποίηση κατάλληλων διεπαφών σημασιολογικής ομαδοποίησης των όρων. Τέλος, θα είχε ενδιαφέρον να διερευνηθεί πώς θα μπορούσε να ενδυναμωθεί η συλλογή όρων από την εφαρμογή μας με σύγχρονους αλγορίθμους μηχανικής μάθησης.

7 Παράρτημα

7.1 Αλγόριθμος Δοκιμής και Αξιολόγησης Συστήματος

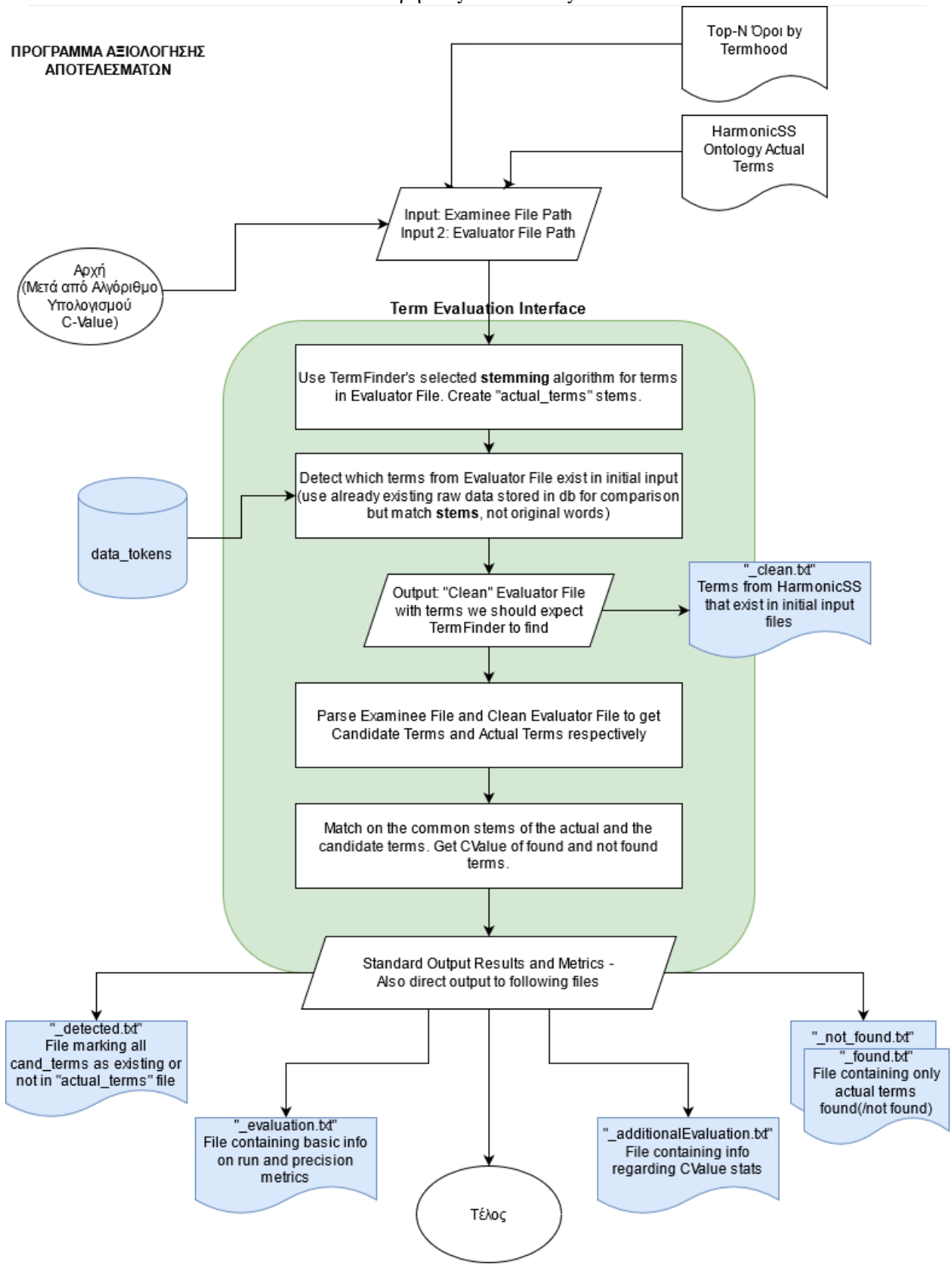
Εδώ, θα αναλύσουμε λίγο παραπάνω τη λειτουργία του Συστήματος Δοκιμής και Αξιολόγησης. Σε πρώτη φάση αναλύουμε τη λειτουργία του και στη συνέχεια δίνουμε και ένα αρχιτεκτονικό διάγραμμα.

Πρώτα, κάνουμε stemming στους όρους που μας δίνονται να ψάξουμε. Μετά, κάνουμε μία αναζήτηση 1 προς 1 των stems των όρων με τα stems που έχουμε εντοπίσει από όλα τα tokens των κειμένων και έχουμε ήδη αποθηκευμένα στον πίνακα data_token. Έτσι, παράγουμε την επονομαζόμενη "clean" οντολογία.

Στη συνέχεια, επανερχόμαστε στο output του *TermFinder*: το αρχείο με τους Top-N κατά C-Value υποψήφιους όρους. Για αυτούς τους όρους, μιας και έχουμε το stem τους, απλά εξετάζουμε αν υπάρχουν ή όχι στην clean οντολογία μας με σύγκριση με τα stems των Υπαρχόντων Όρων της. Αντίστροφα, ελέγχουμε και από την οντολογία μας αν έχουμε βρει όλους τους όρους που θέλουμε. Τέλος, υπολογίζουμε όλες τις απαραίτητες μετρικές, οι οποίες φαίνονται και στο σχήμα που ακολουθεί ως outputs: είναι οι επιπλέον άγνωστοι όροι, οι γνωστοί όροι που βρέθηκαν, οι γνωστοί όροι που δεν βρέθηκαν και 2 αρχεία με βασικές και πιο σύνθετες μετρικές για το precision και recall του αλγορίθμου μας, καθώς και για τα εντοπισμένα CValue για κάθε κατηγορία όρων (γνωστών-άγνωστων) που εντοπίσαμε.

Από αυτά τα δεδομένα, δημιουργήσαμε τους πίνακες που είδαμε και στην ενότητα (0).

Ακολουθεί το σχετικό διάγραμμα:



Σχήμα 12 Πρόγραμμα Δοκιμής και Αξιολόγησης Συστήματος

7.2 Ανεπεξέργαστα Δεδομένα 1: Πείραμα 1

Επισυνάπτουμε τους top-100 by C-Value με φθίνουσα σειρά όρους που έκανε capture η εκτέλεση του αλγόριθμου για σύνθετο μοτίβο, με capturing+ (δηλαδή με αναζήτηση και στη στήλη description των excel). Με γαλάζιο φαίνονται οι «γνωστοί» όροι της οντολογίας, ενώ όλοι οι άλλοι είναι «άγνωστοι»:

reference model term	*unit of measurement*	*eular sjogren 's syndrome*	*normal range of values*	*essdai domains*
medical conditions	*quality of life assessment*	*nervous system*	*normal values*	*sjögren syndrome*
health survey score	*score*	*disease*	*term*	*serum complement*
last follow	*visual analogue scale*	*diagnosis*	*pharmaceutical drugs*	*year of diagnosis*
autoimmune thyroid disease	*short form*	*last fup*	*peripheral nervous system*	*central nervous system*
schirmer 's test	*focus score*	*section f*	*patient questionnaire*	*epworth sleepiness scale*
serum protein test	*rheumatoid factor*	*germinal centers*	*dosage units*	*nervous system disease*
volume	*c reactive protein*	*need of improvement*	*antinuclear antibodies*	*antirheumatic drug*
unit	*medication startdate*	*medication stopdate*	*dry mouth*	*greenspan focus score*
data	*patient*	*value*	*presence*	*raynaud 's phenomenon*
year of birth	*diagnosis year*	*c4 levels*	*urogenital system disease*	*domains*
c3 levels	*dry eyes*	*date*	*mg*	*essdai*
essdaif	*immunoglobulin g*	*worst eye*	*ocular staining score*	*dl normal range of values*
digestive system disease	*years*	*salivary gland biopsy*	*euroqol group*	*conventional disease*
hematologic test	*factor of conversion*	*correspondence with hamonicss*	*essdai domain score*	*year*
mass	*liver disease*	*last visit*	*field*	*disease diagnosis*
rose bengal staining	*profile of fatigue*	*level of activity*	*other autoantibody test*	*hospital anxiety*
level	*minor salivary gland*	*diagnosis date*	*normal limits*	*anno_dg*
monoclonal m	*palpable purpura*	*unstimulated saliva flow value*	*van bijsterveld score*	*eye moisturizing agent*
domain score weight	*free light*	*assessment*	*serum*	*frequency units*

Πίνακας 23: Top 100 Γνωστοί ή Άγνωστοι Όροι – Πείραμα 1

7.3 Ανεπεξέργαστα Δεδομένα 2: Πείραμα 2

Επισυνάπτουμε τους top-100 by C-Value όρους με φθίνουσα σειρά που εντόπισε ο TermFinder για σύνθετο συντακτικό μοτίβο, στην περίπτωση του δεύτερου πειράματος από την περιγραφή Κλινικών Δοκιμών. Με γαλάζιο φαίνονται οι «γνωστοί» όροι της οντολογίας, ενώ όλοι οι άλλοι είναι «άγνωστοι»:

inclusion criteria	*exclusion criteria*	*sjogren 's syndrome*	*salivary glands*	*autoimmune disease*
patients	*systemic lupus erythematosus*	*syndrome*	*study*	*disease*
criteria	*lupus erythematosus*	*dry mouth*	*systemic lupus*	*informed consent*
disease activity	*ocular surface*	*quality of life*	*treatment*	*rheumatic diseases*
dry eye syndrome	*european consensus group*	*primary sjogrens syndrome*	*rheumatoid arthritis*	*clinical trials*
systemic disease	*years of age*	*salivary flow*	*european consensus criteria*	*rheumatoid arthrits*
systemic autoimmune disease	*primary sjogren syndrome*	*european criteria*	*primary sjogren 's syndrome*	*subjects*
clinical trial	*classification criteria*	*week*	*minor salivary glands*	*other diseases*
sjogren	*severe dry eye*	*dry eyes*	*american college of rheumatology*	*glands*
ss	*life quality*	*connective tissue disease*	*dry eye disease*	*patients with sjogren*
national institutes of health clinical center	*inclusion*	*pss patients*	*disease patients*	*treatment of dry eye*
chronic diseases	*weeks*	*disease activity index*	*salivary gland dysfunction*	*hepatitis b*
patients with pss	*lung diseases*	*european consensus group criteria*	*lacrimal glands*	*months*
university hospital	*at screening*	*hepatitis c*	*healthy controls*	*dry eye*
systemic sclerosis	*sjogren patients*	*exclusion*	*participants*	*european league against rheumatism*
connective tissue	*active infection*	*ss patients*	*dry eye patients*	*history*
other autoimmune diseases	*salivary gland biopsy*	*group*	*ocular surface disease*	*health*
childbearing potential	*exocrine glands*	*ocular disease*	*score β*	*tcm tongue diagnosis*
salivary dysfunction	*diagnosis*	*tests*	*oral health*	*screening visit*
symptom	*nervous system*	*host disease*	*patient*	*primary ss*

Πίνακας 24: Top 100 Γνωστοί ή Άγνωστοι Όροι – Πείραμα 2

7.4 Ανεπεξέργαστα Δεδομένα 2: Πείραμα 3

Επισυνάπτουμε τους top-100 by C-Value όρους με φθίνουσα σειρά που εντόπισε ο TermFinder για σύνθετο συντακτικό μοτίβο, στην περίπτωση του τρίτου πειράματος από τα ιατρικά άρθρα του CORD-19. Με γαλάζιο φαίνονται οι «γνωστοί» όροι της οντολογίας, ενώ όλοι οι άλλοι είναι «άγνωστοι»:

bone marrow	*%*	*patients*	*platelet counts*	*γδt cells*
αβt cells	*acute leukemia*	*% of patients*	*platelet count*	*disease*
cells	*t cell*	*study*	*peripheral blood*	*b cell*
bone marrow examination	*peripheral smear*	*plasma cells*	*treatments*	*finding*
bone marrow biopsy	*treatment*	*clinical trials*	*% cases*	*immune response*
bone marrow aspiration	*results*	*lung disease*	*marrow*	*diagnosis*
patient	*retrospective study*	*blood*	*flow cytometry*	*risk factors*
plasma cell	*case reports*	*present study*	*response*	*groups*
interstitial lung disease	*marrow aspirate*	*cases*	*marrow biopsy*	*baseline baseline*
case	*bone biopsy*	*lilra3 d1*	*p value*	*myeloid leukemia*
autoimmune disease	*complete blood counts*	*cell lymphoma*	*acute myeloid leukemia*	*chronic disease*
iron deficiency	*infection*	*serum levels*	*case report*	*levels*
blood counts	*studies*	*lymph nodes*	*years of age*	*% of cases*
medical college	*prospective study*	*systematic review*	*serum ferritin*	*lungs*
counts	*complete blood count*	*methods*	*median age*	*conclusion*
multiple myeloma	*blood transfusions*	*total counts*	*discussion*	*years*
covid	*interstitial disease*	*b cell follicles*	*significant difference*	*study cohort*
lung	*lymph node*	*rheumatic disease*	*therapy*	*pleural effusion*
platelet	*acute lymphoblastic leukemia*	*fi brosis*	*objectives*	*total count*
control group	*s. pneumoniae*	*hodgkin lymphoma*	*count*	*blood samples*

Πίνακας 25: Top 100 Γνωστοί ή Άγνωστοι Όροι – Πείραμα 3

8 Βιβλιογραφία

[1] H. Guo, L. Wang, F. Chen, and D. Liang, "Scientific big data and Digital Earth," *Chinese Science Bulletin (Chinese Version)*, vol. 59, p. 1047, Sep. 2014, doi: 10.1360/972013-1054.

[2] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World from Edge to Core," 2018. [Online]. Available: http://cloudcode.me/media/1014/idc.pdf

[3] "COVID-19 Open Research Dataset Challenge (CORD-19) | Kaggle." https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (accessed Sep. 11, 2021).

[4] "HarmonicSS – HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making." https://www.harmonicss.eu/ (accessed Sep. 18, 2021).

[5] R. Grishman, *Computational linguistics : an introd.* Cambridge Univ. Pr, 1989.

[6] S. Dipper, "Theory-driven and corpus-driven computational linguistics and the use of corpora," 2008.

[7] N. Chomsky, "Three models for the description of language," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113–124, 1956, doi: 10.1109/TIT.1956.1056813.

[8] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011, doi: 10.1136/amiajnl-2011-000464.

[9] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of Biomedical Informatics*, vol. 37, no. 6, pp. 512–526, 2004, doi: 10.1016/j.jbi.2004.08.004.

[10] K. Heylen, D. D. H.-H. of terminology, and undefined 2015, "Automatic term extraction," *books.google.com*, Accessed: Sep. 16, 2021. [Online]. Available: https://www.google.com/books?hl=el&lr=&id=MQZoBwAAQBAJ&oi=fnd&pg=PA203&dq=Automatic+Term+Extraction+Kris+Heylen+and+Dirk+De+Hertog&ots=KpdC2GeHdT&sig=ww7BJtlCNDg2psjRKqwbl41AU9Y

[11] K. Kageura and B. Umino, "Methods of automatic term recognition: A review," *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 3, no. 2, pp. 259–289, 1996, doi: 10.1075/term.3.2.03kag.

[12] H. Nakagawa and T. Mori, "Automatic term recognition based on statistics of compound nouns and their components," *Terminology*, vol. 9, no. 2, pp. 201–219, 2003, doi: 10.1075/term.9.2.04nak.

[13] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the C-value/NC-value method," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000, doi: 10.1007/s007999900023.

[14] I. Spasić, M. Greenwood, A. Preece, N. Francis, and G. Elwyn, "FlexiTerm: a flexible term recognition method," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 27, 2013, doi: 10.1186/2041-1480-4-27.

[15] V. Gayen and K. Sarkar, "A Machine Learning Approach for the Identification of Bengali Noun-Noun Compound Multiword Expressions," *CoRR*, vol. abs/1401.6567, 2014, [Online]. Available: http://arxiv.org/abs/1401.6567

[16] N. J. van Eck, L. Waltman, E. C. M. Noyons, and R. K. Buter, "Automatic term identification for bibliometric mapping," *Scientometrics 2010 82:3*, vol. 82, no. 3, pp. 581–596, Feb. 2010, doi: 10.1007/S11192-010-0173-0.

[17] G. Nenadic, S. Ananiadou, and J. McNaught, "Enhancing automatic term recognition through recognition of variation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, Aug. 2004, pp. 604–610. [Online]. Available: https://aclanthology.org/C04-1087

[18] I. P. and M. S. Korkontzelos Ioannis and Klapaftis, "Reviewing and Evaluating Automatic Term Recognition Techniques," in *Advances in Natural Language Processing*, 2008, pp. 248–259.

[19] D. Fedorenko, N. Astrakhantsev, and D. Turdakov, "Automatic Recognition of Domain-Specific Terms: an Experimental Evaluation," *Proceedings of the Institute for System Programming of RAS*, vol. 26, pp. 55–72, Sep. 2014, doi: 10.15514/ISPRAS-2014-26(4)-5.

[20] M. Granitzer, V. Sabol, K. W. Onn, D. Lukose, and K. Tochtermann, "Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques," *Future Internet*, vol. 2, no. 3, pp. 238–258, 2010, doi: 10.3390/fi2030238.

[21] M. da S. Conrado, T. A. S. Pardo, and S. O. Rezende, "A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set *," pp. 16–23, 2013, Accessed: Sep. 18, 2021. [Online]. Available: http://www.nilc.icmc.usp.br/

[22] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.

[23] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Comput. Linguist.*, vol. 19, no. 2, pp. 313–330, 1993.

[24] "Stop Words List • Key Content." https://key-content.com/stop-words-list/ (accessed Sep. 18, 2021).

[25] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," 2006.