Ιονιο Πανεπιστημιο

Τμήμα Πληροφορικής

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Ακαδ. Έτος: 2021 - 2022

ΕΡΓΑΣΙΑ

ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΘΕΜΑΤΙΚΩΝ ΟΡΩΝ ΣΕ ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ

Στοιχεία Φοιτητών:

Αριθμός μητρώου	Ονοματεπώνυμο	E-mail
Π2018219	**Χρήστος Μήλιος**	**p18mili1@ionio.gr**
Π2018151	**Βασίλειος Γαβριηλίδης**	**p18gavr@ionio.gr**
Π2018121	**Γεώργιος-Σπυρίδων Ρώσσης**	**p18ross@ionio.gr**
Π2018104	**Γεωργία Κερασίδου**	**p18kera@ionio.gr**
Π2017160	**Παναγιώτης Σιώλος**	**p17siol@ionio.gr**
Π2016041	**Ιωάννης Αλεξανδρίδης**	**p16alex@ionio.gr**

Κέρκυρα, Φεβρουάριος 2022

1) Ερευνητικό Πρόβλημα

Αντικείμενο της εργασίας ήταν η αυτόματη αναγνώριση όρων μέσα από σώματα κειμένων. Η αυτόματη αναγνώριση όρων παίζει σημαντικό ρόλο τόσο στην κατανόηση της φυσικής γλώσσας όσο και στην επεξεργασία μηχανικής γλώσσας. Έτσι κάποιος μπορεί πολύ πιο εύκολα και γρήγορα να εντοπίσει το περιεχόμενο ενός κειμένου και συχνά κρίνεται χρήσιμη και στην διαδικασία της μετάφρασης. Οι αλγοριθμοί εστιάζουν στην μέγιστη δυνατή ακρίβεια της ορολογίας που υποστηρίζει την ραγδαία εξέλιξη της τεχνολογίας και της επιστήμης αφού συνεχώς υπάρχει ανάγκη για δημιουργία νέων όρων.

Αρχικά, μελετήσαμε διάφορες έρευνες και εργασίες που είχαν πραγματοποιηθεί γύρω από αναγνώριση όρων για να αποκτήσουμε μία καλύτερη εικόνα του θέματος και να δούμε τεχνικές υλοποίησης που χρησιμοποιήθηκαν. Ενδεικτικά αναφέρονται οι ακόλουθες:

Automatic Term Extraction And Document Similarity in Special Text Corpora, *E. Milios, Y. Zhang, B. He and L. Dong*, σύγκριναν κείμενα τεχνολογίας και ιατρικής παρατηρώντας μεγάλες ομοιότητες.

A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set, *Merley Da Silva Conrado , Thiago A. Salgueiro Pardo , Solange Oliveira Rezende*, ανέλυσαν μόνο unigrams όρους χρησιμοποιώντας Naive Bayes (Probabilistic), JRip (Rule Induction), J48 (Decision Tree) και SMO (Statistical Learning).

An improved Corpus Comparison Approach to Domain Specific Term Recognition, *Xiaoyue Liu, and Chunyu Kit*

Automatic recognition of multi-word terms:. the C-value/NC-value method, *Katerina Frantzi, Sophia Ananiadou & Hideki Mima*, οι οποίοι ανεξάρτητα από το θεματικό τομέα δημιούργησαν αυτόματη αναγνώριση όρων πολλαπλών λέξεων. Συγκεκριμένα, η τελευταία έρευνα αποτελεί σημαντικό σημείο για την επιστήμη της αναγνώρισης όρων, αφού πέρα του ότι ήταν από τις πρώτες έρευνες πάνω σε αυτό το θέμα, δημιουργήθηκε ένας αλγόριθμος βάση που χρησιμοποιήθηκε ως βιβλιογραφία ακόμα και σε μερικές από τις προαναφερόμενες έρευνες.

Τέλος, το **Text categorization with WEKA: A survey - ScienceDirect**, Donatella Merlini, Martina Rossini βασίζεται στην εξόρυξη πληροφοριών από την διάγνωση ενός σετ κειμένων . Βοήθησε στην διαδικασία της προεπεξεργασίας όσο και στην ανάλυση των τελικών αποτελεσμάτων.

Από την συγκεκριμένη βιβλιογραφία διακρίνεται πως οι δημοφιλέστεροι αλγόριθμοι για την ταξινόμηση των όρων είναι οι Bayes classifiers και k-nearest neighbours.

2) Σετ Δεδομένων

Για τα δεδομένα ειδικού σκοπού χρησιμοποιήθηκαν 6 κείμενα όλα σχετικά με τον τομέα της πληροφορικής, καθώς έχουν περιεχόμενο άμεσο με τον τομέα που διαλέξαμε. Δεν αποκλείστηκε κάποια κατεύθυνση της επιστήμης και εστιάσαμε ισορροπημένα σε θέματα software και hardware.

Συγκεκριμένα χρησιμοποιήθηκαν τα παρακάτω έγγραφα ή αποσπάσματα αυτών:

- **The Tera computer system** *Robert Alverson, David Callahan, Daniel Cummings, Brian Koblenz, Allan Porterfield, Burton Smith*, από το οποίο χρησιμοποιήθηκαν 2150 λέξεις.
- **Deep Learning for Computer Vision: A Brief Review** *Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis and Eftychios Protopapadakis*, άρθρο από το οποίο χρησιμοποιήθηκαν 4900 λέξεις.
- **The role of psychology in understanding the impact of computer games** *Elizabeth Boyle, Thomas M.Connolly, Thomas Hainey*. Από την έρευνα αυτή χρησιμοποιήθηκαν 3550 λέξεις.
- **Algorithms and Complexity** *Emanuele Viola*, απ' όπου αντλήθηκαν περίπου 1350 λέξεις.
- **An Abstract View of Programming Languages** *Eugenio Moggi*, το οποίο έδωσε περίπου 4750 λέξεις.
- **Computer Architecture** *Michael Flynn*. Από το βιβλίο αυτό χρησιμοποιήθηκαν γύρω στις 7000 λεξεις.

Όλα τα κείμενα ειδικού σκοπού ήταν περίπου 24.000 λέξεις στο σύνολο και κρατήσαμε τις μοναδικές λέξεις όπου τελικά ήταν 5440. Στου γενικού σκοπού επιλέξαμε κείμενα από ειδήσεις του 2018. Απο μια πληθώρα θεματικών άρθρων. Το σύνολο των λέξεων ήταν περίπου 40.000 λέξεις.

3) Επιλογή Χαρακτηριστικών

Η επιλογή των χαρακτηριστικών έγινε κυρίως με γνώμονα τις συνήθεις πρακτικές που ακολουθούνται και προτείνονται από τη βιβλιογραφία. Τα περισσότερα από αυτά είναι στατιστικά (statistical), δύο είναι γλωσσολογικά (linguistic) και ένα είναι μικτό υπό την έννοια ότι μπορεί να ερμηνευθεί με διαφορετικό πρίσμα.

Στατιστικά χαρακτηριστικά

Ένα από τα βασικά στατιστικά χαρακτηριστικά είναι η Σχετική Συχνότητα Εμφάνισης της Λέξης. Στο γενικό σώμα κειμένων (ΓΣΚ) εκφράζεται από το λόγο του αριθμού εμφάνισης του πιθανού όρου/λέξης σε όλα τα κείμενα προς το συνολικό αριθμό λέξεων όλων του σώματος κειμένων. Στο σετ δεδομένων του ΓΣΚ, όσο μικρότερος είναι αυτός ο λόγος για μία λέξη, τόσο πιο πιθανό είναι αυτή η λέξη να είναι όρος, ενώ αντίθετα όσο ο λόγος αυτός μεγαλώνει τόσο αυξάνεται η πιθανότητα η λέξη να μην είναι όρος. Το επόμενο χαρακτηριστικό, η Σχετική Συχνότητα Εμφάνισης της Λέξης στο ειδικό σώμα κειμένων (ΕΣΚ), παρά το γεγονός ότι εκφράζεται όπως προηγουμένως από το λόγο του αριθμού εμφάνισης του πιθανού όρου/λέξης προς το συνολικό αριθμό των λέξεων, η ερμηνεία είναι διαφορετική αφού τώρα όσο μεγαλύτερος είναι αυτός ο λόγος για μία λέξη, τόσο πιο πιθανό είναι η λέξη αυτή να είναι όρος και αντίστοιχα όσο ο λόγος αυτός μικραίνει τόσο αυξάνεται η πιθανότητα η λέξη να μην είναι όρος.

Ένα ακόμη χαρακτηριστικό είναι το TF (Term Frequency), που μας δείχνει το πόσο συχνά εμφανίζεται ο όρος σε κάθε έγγραφο. Το χαρακτηριστικό αυτό προσφέρει χρησιμότερη πληροφορία για έγγραφα με μεγάλο αριθμό λέξεων αλλά κυρίως για τον υπολογισμό του γινομένου TF*IDF, όπως επίσης

και το επόμενο χαρακτηριστικό, το DF (Document Frequency), που εκφράζει τον αριθμό των εγγράφων στα οποία εμφανίζεται, έστω και μια φορά, η λέξη. Όσο μεγαλύτερο είναι το DF, τόσο πιο πιθανό να είναι σχετική η λέξη. Ο συνδυασμός αυτών των δύο χαρακτηριστικών, όπως ήδη αναφέραμε, δίνει ένα ακόμα χαρακτηριστικό το γινόμενο TF*IDF (δλδ., το γινόμενο των TF και IDF -inverse DF, όπου IDF είναι ο δεκαδικός λογάριθμος του συνόλου των εγγράφων προς τον αριθμό των εγγράφων που εμφανίζεται ο όρος), το οποίο μας δείχνει το πόσο σημαντική/σχετική είναι η λέξη. Όσο υψηλότερο το αποτέλεσμα τόσο πιο σχετική είναι η λέξη ενώ αντίθετα όσο το αποτέλεσμα πλησιάζει προς το μηδέν η λέξη τείνει να είναι όλο και λιγότερο σχετική.

Τέλος, ένα ακόμη χαρακτηριστικό είναι η Εντροπία της λέξης, η οποία εκφράζει την αβεβαιότητα εμφανισης της πληροφορίας/λέξης που επεξεργαζόμαστε. Όσο πιο υψηλή είναι η τιμή της εντροπίας για μία λέξη τόσο πιο μεγάλη είναι η αβεβαιότητα εμφάνισής της και κατα συνέπεια καθιστά πιο δύσκολη μια σωστή πρόβλεψη.

Γλωσσολογικά χαρακτηριστικά

Τα γλωσσολογικά χαρακτηριστικά είναι η Λέξη και Μέρος του Λόγου (POS), εκφράζει το προφανές, δηλαδή τι μέρος του λόγου είναι η κάθε λέξη όπως έχει επισημανθεί μέσω εφαρμογής της Python με τη χρήση της κατάλληλης βιβλιοθήκης **nltk**.

Χαρακτηριστικό	**Τύπος Υπολογισμού**	**Χαρακτηριστικό**	**Τύπος Υπολογισμού**
Σχετική Συχνότητα Εμφάνισης στο ΕΣΚ	$\frac{\sum \text{εμφανίσεις λέξης}}{\sum \text{λέξεις ΕΣΚ}}$	Σχετική Συχνότητα Εμφάνισης στο ΓΣΚ	$\frac{\sum \text{εμφανίσεις λέξης}}{\sum \text{λέξεις ΓΣΚ}}$
Term Frequency (TF)	$\frac{\sum \text{εμφανίσεις όρου}}{\sum \text{λέξεις εγγράφου}}$	Document Frequency (DF)	$DF = \sum D_{term}$ όπου $term \in D$
TF*IDF	$TF * IDF = TF * \log_{10} \frac{N_D}{DF}$	Μέρος του Λόγου	Επισημειωμένο από τον nltk
Εντροπία	$H(X) = -P(X=0)\log_2 P(X=0) - P(X=1)\log_2 P(X=1)$ όπου X είναι η λέξη και X=0 όταν η λέξη δεν εμφανίζεται και X=1 όταν η λέξη εμφανίζεται		-

Πίνακας 1. Τύποι υπολογισμού χαρακτηριστικών

4) Υλοποίηση Ειδικού Λεξικού και Αποτελέσματα Εκπαίδευσης

Πρώτο Μέρος

Για την υλοποίηση της εργασίας επιλέξαμε να χρησιμοποιήσουμε την γλώσσα Python, να επεξεργαστούμε τις λέξεις των κειμένων στο Excel και τέλος, να εκπαιδεύσουμε και να δοκιμάσουμε διάφορα μοντέλα εκπαίδευσης με το πρόγραμμα Weka. Αρχικά συντάχθηκε κώδικας στην γλώσσα Python με τον οποίο το σύνολο των λέξεων του ειδικού κειμένου πήραν ετικέτα ανάλογα με το μέρος του λόγου στο οποίο ανήκουν. Στη συνέχεια, το κείμενο καθαρίστηκε από σημεία στίξης και λέξεις δίχως σημασιολογική σημασία (stopwords) όπως άρθρα, αντωνυμίες, κτλ, χρησιμοποιώντας ένα σώμα λέξεων που δημιουργήθηκε ειδικά για αυτό τον σκοπό. Έπειτα, όλοι οι κεφαλαίοι χαρακτήρες που υπήρχαν στις λέξεις μετατράπηκαν σε πεζούς και τέλος δημιουργήθηκαν αρχεία τύπου CSV με το σύνολο των "καθαρών λέξεων", στα οποία κάθε λέξη τοποθετήθηκε σε μία γραμμή η μία κάτω από την άλλη. Ο κώδικας μπορεί να βρεθεί στο συμπληρωματικό υλικό της εργασίας.

Στη συνέχεια, έγινε αυτόματα ο υπολογισμός των χαρακτηριστικών που επιλέχθηκαν, όπως περιγράφεται στην τρίτη ενότητα, και ακολούθως το σύνολο των λέξεων του ειδικού κειμένου επισημειώθηκε με τις ετικέτες "Y" και "N" από τα μέλη της ομάδας. Για την παραπάνω διαδικασία επιλέξαμε να χωριστούμε σε τρεις ομάδες των δύο ατόμων και να χωρίσουμε το σύνολο των λέξεων σε τρία μέρη όπου κάθε λέξη που υπήρχε στα τρία επιμέρους σύνολα πήρε ετικέτα από δύο μέλη της ομάδας. Στη περίπτωση που κάποια λέξη έπαιρνε "Y" από τον ένα επισημειωτή και "N" από τον άλλο τότε εξ' ορισμού δινόταν η ετικέτα "Y" στην λέξη.

Με την παραπάνω διαδικασία δημιουργήθηκε το εκπαιδευμένο σώμα ειδικού κειμένου το οποίο χρησιμοποιήθηκε στο Weka. Στο Weka, στην επιλογή Explorer➜Preprocess, φορτώθηκε το αρχείο και έπειτα στην επιλογή Classify επιλέχθηκε η τεχνική cross validation με Folds=10. Οι αλγόριθμοι που επιλέχθηκαν για τις διάφορες δοκιμές είναι αυτοί που αναφέρονταν συχνά στην βιβλιογραφία, ή παραλλαγές αυτών, ενώ επίσης προστέθηκε το πακέτο Auto-WEKA (από την αρχική διεπαφή GUI➜Tools➜Package Manager➜Auto-WEKA➜Install) το οποίο όταν "τρέχει" υπολογίζει σύμφωνα με τα δεδομένα που του δίνουμε ποιος είναι ο καταλληλότερος αλγόριθμος ή συνδυασμός αλγορίθμων για την εκπαίδευση του σετ μας, με συγκεκριμένες τιμές για τις παραμέτρους διαμόρφωσης (configuration arguments). Ο Πίνακας 2 παραθέτει τους αλγόριθμους που επιλέχθηκαν για εκπαίδευση.

● IBK με KNN = 1,4,8	● Random Forest ATRB[1,2,3,4,19]*
● Bagging (δίχως παραμετροποίηση)	● Logistic Regression
● Bagging (με παραμετροποίηση όπως προτάθηκε από το Auto-WEKA)	● Naive Bayes
● Random Tree	● J48
● Random Forest	● J48 Consolidated

* ATRB: 1 = λέξεις, 2 = σχετική συχνότητα εμφάνισης της λέξης στο ΕΣΚ, 3 = σχετική συχνότητα εμφάνισης της λέξης στο ΓΣΚ, 4 = εντροπία, 19 = είναι όρος? Y/N

Πίνακας 2. Λίστα Αλγορίθμων

Δεύτερο Μέρος

Οι λέξεις γενικού σκοπού επεξεργάστηκαν με τον ίδιο ακριβώς τρόπο όπως και οι λέξεις ειδικού σκοπού με μοναδική διαφορά πως για την επισημείωση της λέξης ως ειδικός όρος συμπληρώθηκε όλη η τελευταία στήλη "είναι όρος Y/N" με το σύμβολο "?" με τελικό στόχο να πάρει ετικέτα βασισμένη στο εκπαιδευμένο μοντέλο που υλοποιήθηκε με την βοήθεια του ειδικού, επισημειωμένου με το χέρι, σώματος κειμένου, όπως αυτό περιγράφηκε παραπάνω. Το συγκεκριμένο σώμα αποτελεί το "άγνωστο" σύνολο λέξεων που πρέπει να πάρει ετικέτα και ονομάζεται σε πολλές εργασίες ως "test set". Το επισημειωμένο, "γνωστό", εκπαιδευμένο σύνολο λέξεων ονομάζεται "training set". Η διαδικασία είναι η εξής: Φορτώνεται το ειδικό σώμα λέξεων (training set) και εκπαιδεύεται με κάποιον από τους αλγόριθμους που αναφέρθηκαν. Έπειτα από το πεδίο "supplied test set" επιλέγεται το αρχείο με τις λέξεις που χρειάζονται επισημείωση (test set). Το αποτέλεσμα του κάθε αλγορίθμου αποθηκεύεται σε ένα αρχείο CSV για περαιτέρω επεξεργασία και την εξαγωγή συμπερασμάτων σχετικά με τα αποτελέσματα της επισημείωσης. Στο παρακάτω σύνδεσμο, υπάρχουν τα αποτελέσματα από την εκπαίδευση του μοντέλου για κάθε αλγόριθμο: Αποτελέσματα εκπαίδευσης - Detailed Accuracy By Class - Confusion Matrix

Τρίτο μέρος

Μία ακόμη μέθοδος που βρήκαμε στην βιβλιογραφία είναι ο διαχωρισμός του εκπαιδευμένου σώματος κειμένου με τους ειδικούς θεματικούς όρους σε σώμα εκπαίδευσης (training set) και σώμα δοκιμής (test set). Για αυτό τον λόγο, δοκιμάσουμε να χωρίσουμε το επισημειωμένο σώμα κείμενο και να χρησιμοποιηθούν οι πρώτες 300 λέξεις (ένα split της τάξης του 6%) για να δοκιμαστεί η απόδοση των διαφόρων μοντέλων. Το θετικό με αυτή την μέθοδο, η οποία δεν αποτελεί την ιδανική, είναι ότι τα συμπεράσματα για το accuracy υπολογίζονται αυτόματα από το Weka. Λόγω περιορισμών στην μνήμη δεν ήταν δυνατόν να πειραματιστούμε με μεγαλύτερο αριθμό λέξεων. Στον παρακάτω σύνδεσμο υπάρχουν τα αποτελέσματα από την δοκιμή διάφορων αλγορίθμων με test set από το ειδικό σώμα κειμένων: Αποτελέσματα από test set του ειδικού κειμένου

5) Αξιολόγηση Αποτελεσμάτων

Όπως είναι λογικό, από την στιγμή που το σώμα λέξεων γενικού σκοπού δεν έχει κάποια ετικέτα από πριν, δεν είναι δυνατόν να εξαχθούν συμπεράσματα σχετικά με το "Detailed Accuracy By Class". Για αυτό τον λόγο, πραγματοποιήθηκε περαιτέρω ανάλυση των αποτελεσμάτων (predictions). Ειδικότερα, χρησιμοποιήθηκε Excel και κώδικας σε python ώστε να βρεθούν πόσες λέξεις από αυτές που επέστρεψαν οι αλγόριθμοι (predictions) είναι πραγματικοί όροι και υπάρχουν στο ειδικό σώμα λέξεων με ετικέτα "Y". Δηλαδή, μετρήθηκαν τα αληθώς θετικά και τα ψευδώς θετικά με τα οποία μπορεί να υπολογιστεί και η ακρίβεια (precision) για τον κάθε αλγόριθμο. Τα αποτελέσματα απεικονίζονται στον Πίνακα 3 ενώ ο κώδικας μπορεί να βρεθεί στο συμπληρωματικό υλικό της εργασίας.

Αντίθετα, η χρήση του ΕΣΚ ως σετ training/testing με split 6% δίνει τη δυνατότητα αυτόματου υπολογισμού από το WEKA των τιμών precision (P) και recall (R). Ο περιορισμός βέβαια του αριθμού των λέξεων (λόγω θεμάτων με τη μνήμη) υπάρχει αλλά μπορούμε να δούμε ότι τα αποτελέσματα που φαίνονται στον Πίνακα 4 είναι αρκετά αξιόλογα για περαιτέρω διερεύνηση.

PREDICTIONS FROM (Ξεχωριστά αρχεία training/test, ΕΣΚ/ΓΣΚ)	TOTAL TRUE TERMS	TRUE POSITIVE	FALSE POSITIVE	PRECISION(%)
IBK KNN8	36	7	29	**19,4%**
IBK KNN4	215	64	151	**29,7%**
IBK KNN1	532	219	313	**41,1%**
BAGGING (χωρίς παρ/ση)	289	288	1	**99,6%**
BAGGING (παραμετροποίηση Auto-WEKA)	6	3	3	**50,0%**
RANDOM TREE	420	291	129	**69,2%**
RANDOM FOREST	65	64	1	**98,4%**
LOGISTIC REGRESSION	2276	287	1989	**12,6%**
NAIVE BAYES	93	22	71	**23,6%**
J48 CONSOLIDATED	6999	102	6897	**1,4%**
RANDOM FOREST ATRB[1,2,3,4,19]	322	321	1	**99,6%**

* ATRB: 1 = λέξεις, 2 = σχετική συχνότητα εμφάνισης της λέξης στο ΕΣΚ, 3 = σχετική συχνότητα εμφάνισης της λέξης στο ΓΣΚ, 4 = εντροπία, 19 = είναι όρος? Y/N

Πίνακας 3. Αξιολόγηση αποτελεσμάτων για διαφορετικά αρχεία train/test

PREDICTIONS FROM (Ιδιο αρχείο training/test, ΕΣΚ split 6%)	CORRECTLY CLASSIFIED INSTANCES (total of 299)	TRUE POS	FALSE POS	TRUE NEG	FALSE NEG	PRECISION	RECALL
IBK KNN4	252	27	19	225	28	0.587	0.491
IBK KNN1	288	48	4	240	7	0.923	0.873
RANDOM TREE	245	1	0	244	54	1.000	0.018
LOGISTIC REGRESSION	252	14	6	238	41	0.700	0.255
NAIVE BAYES	234	31	41	203	24	0.431	0.564
J48	258	15	1	243	40	0.938	0.273

Πίνακας 4. Αξιολόγηση αποτελεσμάτων για ίδιο αρχείο train/test (splitting)

Όπως φαίνεται από τον Πίνακα 3, ο Simple Bagging, ο Random Forest με όλα τα χαρακτηριστικά και ο Random Forest με 5 χαρακτηριστικά παρουσίασαν πάρα πολύ υψηλό ποσοστό ακρίβειας, 98.4% ο δεύτερος και 99.6% οι άλλοι δύο και μάλιστα οι τελευταίοι με μεγάλο αριθμό όρων, περίπου το μισό του συνόλου των όρων του ΕΣΚ. Ο Random Tree παρουσίασε αξιοπρεπή αποτελέσματα της τάξης του 70% σχεδόν ενώ οι υπόλοιποι απέτυχαν να δώσουν αξιοποιήσιμα αποτέλεσμα με εξαίρεση ίσως τον ο $IBK_{KNN=1}$ που σε συνδυασμό με επισημειωμένο test set (για τον υπολογισμό σωστού recall με τις TRUE και FALSE αρνητικές προβλέψεις) να πετύχει καλύτερα αποτελέσματα. Τέλος, ο παραμετροποιημένος Bagging που πρότεινε το WEKA ναι μεν παρουσίασε ακρίβεια 50% αλλά αναγνώρισε μόνο 6 λέξεις ως όρους.

Υπήρχε η σκέψη να κάνουμε μία καταγραφή των όρων που υπάρχουν στο ΓΣΚ για να μπορέσουμε να δούμε και την απόδοση των αλγορίθμων ως προς την ανάκληση (recall) αλλά η τιμή αυτή δεν θα ήταν καθόλου αντιπροσωπευτική αφού απλά θα είχαμε γνώση, για παράδειγμα, για το πόσοι όροι που υπάρχουν στο ΓΣΚ δεν έχουν αναγνωριστεί ως όροι αλλά όχι πόσες φορές ο καθένας και αν καλώς δεν αναγνωρίστηκαν ως όροι ή όχι. Αυτό για να γίνει θα έπρεπε να επισημειωθεί όλο το ΓΣΚ λέξη κατά λέξη και γι' αυτό το λόγο δεν έχουμε υπολογίσει τιμή για την ανάκληση.

Από τον Πίνακα 4, μπορεί να δει κανείς ότι οι περισσότεροι αλγόριθμοι επέδειξαν άνω του μετρίου τιμές ακρίβειας αλλά γενικά δεν τα πήγαν καλά με την ανάκληση. Εξαίρεση αποτελεί ο IBK KNN=1, ο έκανε σωστή ταξινόμηση για τους 288 από τους 299 όρους με ακρίβεια και ανάκληση αντίστοιχα 92.3% και 87.3%.

6) Συμπεράσματα - Σύγκριση

Η αυτοματοποιημένη αναγνώριση θεματικών όρων είναι ένα ιδιαίτερα ενδιαφέρον και απαιτητικό κομμάτι της επιστήμης της Γλωσσικής Τεχνολογίας. Η παρούσα εργασία αποτυπώνει μια προσπάθεια αυτόματης εξαγωγής όρων από ένα συγκεκριμένο επιστημονικό πεδίο στηριζόμενη σε ήδη γνωστά χαρακτηριστικά και αλγόριθμους χωρίς να επιχειρεί να δημιουργήσει κάποιο δικό της μοντέλο. Η βασική μέθοδος για την εκπαίδευση ήταν η χρήση ξεχωριστού αρχείου, δηλαδή του ΕΣΚ που προέκυψε μετά την προ-επεξεργασία, και στη συνέχεια με χρήση του ΓΣΚ ως test set. Εκ του αποτελέσματος μπορούμε να πούμε ότι η δομή των σωμάτων ήταν τέτοια (η επιλογή των ΕΣΚ και ΓΣΚ έγινε τυχαία) που ανέδειξε περισσότερο τους αλγόριθμους ensemble, όπως ο Bagging και ο Random Forest, βασιζόμενοι τόσο στην κατανομή των όρων, όσο και στη συχνότητα εμφάνισης μέσα στο σώμα εμφανίζοντας σχεδόν απόλυτα ποσοστά όσον αφορά την ακρίβεια στην ανάκτηση των όρων.

Με την ολοκλήρωση της εργασίας καταλήξαμε σε ορισμένα συμπεράσματα από την μελέτη της βιβλιογραφίας αλλά και από την δική μας εμπειρία κατά την υλοποίηση και ολοκλήρωση της συγκεκριμένης προσπάθειας. Όπως αναφέρεται και στην βιβλιογραφία αλλά και όπως διαπιστώσαμε οι ίδιοι, για την εξαγωγή ορθών αποτελεσμάτων και την επαρκή εκπαίδευση ενός μοντέλου χρειάζεται μεγάλος αριθμός θεματικών, επισημειωμένων κειμένων και εντατική μελέτη για την επιλογή των χαρακτηριστικών. Όπως είναι λογικό, για την επεξεργασία αρχείων μεγάλου μεγέθους είναι απαραίτητη και μεγάλη υπολογιστική δύναμη. Καταλήγουμε λοιπόν, στα δύο πρώτα σημεία τα οποία χρήζουν βελτίωση στην συγκεκριμένη εργασία, δηλαδή στο μέγεθος του ειδικού λεξιλογίου και στην υπολογιστική δύναμη που είχαμε στην διάθεση μας. Επίσης, για την καλύτερη εξαγωγή συμπερασμάτων σχετικά με τα αποτελέσματα των αλγορίθμων είναι ωφέλιμο να υπολογιστούν και οι τιμές των accuracy, recall διότι προσφέρουν επιπλέον πληροφορίες σχετικά με την απόδοση του κάθε μοντέλου.

7) Βιβλιογραφία

[1] Conrado, M.D., Pardo, T.A., & Rezende, S.O. (2013). A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. *NAACL*.

[2] Kováříková,D.(2021).Machine Learning in Terminology Extraction from Czech and English Texts. Linguistic Frontiers,0(0) -. https://doi.org/10.2478/lf-2021-0001

[3] Milios, E.E., Zhang, Y., & Dong, L. (2003). Automatic term extraction and document similarity in special text corpora.

[4] Liu, X., & Kit, C. (2008). An improved corpus comparison approach to domain specific term recognition. 253-261. Paper presented at 22nd Pacific Asia Conference on Language, Information and Computation, PACLIC 22, Cebu, Philippines.

[5] Alverson, R., Callahan, D., Cummings, D., Koblenz, B.D., Porterfield, A., & Smith, B.J. (1990). The Tera computer system. *ICS 1990*.

[6] Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience : CIN, 2018, 13.

[7] Boyle, E.A., Connolly, T.M., & Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertain. Comput., 2*, 69-74.

[8] Flynn, M. (2007). Computer Architecture. In Wiley Encyclopedia of Computer Science and Engineering, B.W. Wah (Ed.).

[9] Eugenio M. (1989). An Abstract View of Programming Languages. University of Edinburgh EH9 3JZ Edinburgh

[10] Emanuele V. (2019). Algorithms and Complexity

[11] Automatic recognition of multi-word terms:. the C-value/NC-value method, Katerina Frantzi, Sophia Ananiadou & Hideki Mima

[12] Text categorization with WEKA: A survey - ScienceDirect, Donatella Merlini, Martina Rossini