



ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΣΥΝΤΑΚΤΙΚΗΣ ΓΝΩΣΗΣ ΑΠΟ ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ ΤΟΥ ΚΟΙΝΩΝΙΚΟΥ ΙΣΤΟΥ

(SYNTACTIC DATA MINING FROM SOCIAL WEB CORPORA)

ΔΕΣΠΟΙΝΑ ΣΙΔΗΡΟΠΟΥΛΟΥ (Α.Μ. Π2013009)

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ

ΚΕΡΜΑΝΙΔΟΥ ΚΑΤΙΑ

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

ΜΥΛΩΝΑΣ ΦΟΙΒΟΣ-ΑΠΟΣΤΟΛΟΣ & ΣΤΕΦΑΝΙΔΑΚΗΣ ΜΙΧΑΛΗΣ

ΚΕΡΚΥΡΑ 2019

ΕΥΧΑΡΙΣΤΙΕΣ

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά την επιβλέποντα καθηγήτρια της πτυχιακής μου εργασίας κα Κάτια Κερμανίδου επειδή μου έδωσε την ευκαιρία που τόσο ήθελα για την ενασχόλησή μου με την τεχνητή νοημοσύνη γενικότερα. Η καθοδήγηση της σε αυτή τη διαδρομή και ο χρόνος που αφιέρωσε ήταν πραγματικά πολύτιμα για μένα.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου που στάθηκαν στο πλευρό μου και με εμπύχωσαν για την ολοκλήρωση του στόχου μου με αγάπη και δύναμη.

Τέλος, να ευχαριστήσω όλους τους καθηγητές του τμήματος και ιδιαίτερα τους καθηγητές Φοίβο Μυλωνά και Μιχαήλ Στεφανιδάκη που δέχτηκαν να είναι μέλη της τριμελούς επιτροπής αξιολόγησης της πτυχιακής μου εργασίας.

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία εκπονήθηκε στο πλαίσιο απόκτησης πτυχίου του Τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου. Αποτελεί το τελευταίο μέρος για την ολοκλήρωση απόκτησης του πτυχίου.

Αφορά την εξόρυξη συντακτικής γνώσης από σώματα κειμένων του κοινωνικού ιστού. Το σώμα κειμένων που θα χρησιμοποιηθεί, έχει συλλεχθεί από το twitter. Αποτελείται από κάποιες εκατοντάδες tweets, τα οποία συλλέχθηκαν με τη βοήθεια κάποιων γραμμών κώδικα.

Η εξόρυξη της συντακτικής γνώσης υλοποιείται με τεχνικές της μηχανικής μάθησης. Χρησιμοποιούνται χίλια παραδείγματα σε αρχείο excel, για να γίνουν κάποια πειράματα μηχανικής μάθησης. Με αυτό τον τρόπο εκπαιδεύεται ένας αλγόριθμος ταξινόμησης λέξεων σε ουσιαστικό, αντικείμενο ή σε κανένα από τα δύο.

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

Εισαγωγή-Σκοπός

Κεφάλαιο 1 – Σώματα κειμένων και χαρακτηριστικά

1.1 Είδη σωμάτων κειμένων

1.1.1 Εύρος γλωσσικής κάλυψης

1.1.2 Επιλογή γλώσσας

1.1.3 Χρονική κάλυψη

1.1.4 Τροπικότητα

1.1.5 Επισημείωση

1.1.6 Ανανεωσιμότητα

1.2 Σχεδιασμός σωμάτων κειμένων

1.2.1 Περιεχόμενο

1.2.2 Μέγεθος

1.2.3 Τήρηση μεταδεδομένων

1.2.4 Δειγματοληψία

1.2.5 Αντιπροσωπευτικότητα

1.2.6 Διαθεσιμότητα

1.3 Συγκρότηση σωμάτων κειμένων

Κεφάλαιο 2 – Τα γλωσσικά επίπεδα

2.1 Φωνολογικό επίπεδο

2.2 Μορφολογικό επίπεδο

2.3 Συντακτικό επίπεδο

2.4 Σημασιολογικό επίπεδο

2.5 Πραγματολογικό επίπεδο

Κεφάλαιο 3 – Προσεγγίσεις συντακτικής ανάλυσης

3.1 Συντακτική ανάλυση

3.1.1 Τα βασικά μέρη της πρότασης

3.1.2 Ονοματική και ρηματική φράση

3.1.3 Συντακτικές κατηγορίες

3.2 Αλγόριθμοι ταξινόμησης

3.2.1 Επιβλεπόμενη μάθηση

3.2.2 Μη επιβλεπόμενη μάθηση

Κεφάλαιο 4 – Αναγνώριση συντακτικών συσχετίσεων σε αγγλικά tweets

4.1 Συγκέντρωση αγγλικών tweets

4.2 Χαρακτηριστικά μηχανικής μάθησης

4.3 Αποτελέσματα πειραμάτων

4.4 Συμπεράσματα πειραμάτων

Βιβλιογραφία

ΕΙΣΑΓΩΓΗ

Η παρούσα πτυχιακή εργασία έχει ως κύριο στόχο την εξόρυξη συντακτικής γνώσης από σώματα κειμένων του κοινωνικού ιστού, γραμμένα στην αγγλική γλώσσα. Δευτερεύων στόχος της είναι η συγκρότηση ενός αγγλικού σώματος κειμένων, από το μέσο κοινωνικής δικτύωσης που λέγεται Twitter. Για την ανάλυση των δεδομένων της εργασίας και την εξαγωγή αποτελεσμάτων, θα χρησιμοποιηθεί το εργαλείο RapidMiner.

Ως εξόρυξη συντακτικής γνώσης ορίζεται η διαδικασία της χρήσης μιας ή περισσότερων τεχνικών μηχανικής μάθησης υπολογιστών για την αυτόματη εξαγωγή γνώσης συντακτικού περιεχομένου από κάποια δεδομένα. Θεωρείται μια διαδικασία εξαγωγής χρήσιμων πληροφοριών, καθώς είναι χρήσιμη και στην ανάλυση δεδομένων, την ταξινόμηση και τον συνδυασμό των σχέσεων που προκύπτουν.

Η εξόρυξη γνώσης είναι η διαδικασία σύνδεσης των δύο ξεχωριστών συστημάτων της πληροφοριακής τεχνολογίας, της συναλλαγής και της ανάλυσης. Το λογισμικό της αναλύει τις σχέσεις και τα μοτίβα των αποθηκευμένων δεδομένων από τις βάσεις δεδομένων.

Η διαδικασία της εξόρυξης γνώσης περιλαμβάνει ως πρώτο βήμα την συλλογή των δεδομένων και τον καθαρισμό τους. Στη συνέχεια, γίνεται επιλογή των σημαντικών γνωρισμάτων και η εφαρμογή των αλγορίθμων για την εξόρυξη δεδομένων. Τέλος, η διαδικασία ολοκληρώνεται με την απεικόνιση και την αξιολόγηση των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 1

ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

1.1 Είδη σωμάτων κειμένων

Ο πιο εύκολος τρόπος για να φτιάξει κάποιος μια γλωσσική ανάλυση τηρώντας τη μεθοδολογία σωμάτων κειμένων είναι να πάρει και να χρησιμοποιήσει ένα σώμα κειμένων το οποίο έχει φτιαχτεί ήδη από άλλους. Για ορισμένες ερωτήσεις μπορεί να μην υπάρχει στη διάθεσή μας ένα συγκεκριμένο σώμα κειμένων. Οι ιδιαιτερότητες που έχει ένα συγκεκριμένο υλικό κάνουν αντανάκλαση στις έρευνες τα ευρήματα και επηρεάζουν τα συμπεράσματα που κάθε φορά εξάγονται από αυτήν. Για αυτό το λόγο, πρέπει να αντιστοιχίζεται η γλωσσική ποικιλία που εξετάζεται με το γλωσσικό υλικό που συλλέγεται. Έτσι, δεν μπορεί να γίνει ανάλυση στη γλώσσα της τηλεφωνικής συνδιάλεξης συλλέγοντας δεδομένα από μέσα μαζικής ενημέρωσης ή να μελετηθεί ο πολιτικός λόγος συγκεντρώνοντας προφορικό λόγο από τον κοινωνικό περίγυρο. Άρα, είναι άκρως απαραίτητο να γνωρίζουμε πώς συγκροτούνται τα σώματα κειμένων και τις βασικές αρχές σχεδιασμού τους είτε έχουμε ένα σώμα κειμένων που έχει ήδη συγκροτηθεί είτε έχουμε αποφασίσει να φτιάξουμε ένα δικό μας σώμα κειμένων για να μπορέσουμε να εξασφαλίσουμε ότι υπάρχει αντιστοιχία ανάμεσα στη γλωσσική ποικιλία και το γλωσσικό υλικό.

Για αυτό λοιπόν, πριν από τη δημιουργία ενός σώματος θα πρέπει να ξέρουμε σε ποιο είδος σώματος αποσκοπούμε. Τα είδη σωμάτων διακρίνονται από κάποια χαρακτηριστικά τα οποία είναι τα παρακάτω.

1.1.1 Εύρος γλωσσικής κάλυψης

Ανάλογα με το εύρος των γλωσσικών ποικιλιών που καλύπτουν τα σώματα κειμένων χωρίζονται σε γενικά σώματα κειμένων και σε ειδικά σώματα κειμένων. Τα πρώτα αποτελούνται από ένα ευρύ φάσμα ειδών κειμένων και από ποικίλες πηγές και έχουν σαν στόχο να μας δώσουν στοιχεία για μια γλώσσα σε όλο της το εύρος. Τα ειδικά σώματα κειμένων αναφέρονται σε μια από τις παραμέτρους της γλωσσικής ποικιλότητας.

1.1.2 Επιλογή γλώσσας

Το πρώτο βασικό χαρακτηριστικό που έχει ένα σώμα κειμένων είναι η γλώσσα στην οποία βρίσκεται το γλωσσικό υλικό που περιέχει. Λαμβάνοντας υπόψη αυτό το χαρακτηριστικό τα σώματα χωρίζονται σε μονόγλωσσα, αν τα κείμενά τους είναι σε μία μόνο γλώσσα, και πολύγλωσσα, αν είναι σε περισσότερες από μία γλώσσες. Τα πρώτα περιλαμβάνουν λέξεις, προτάσεις ή και μικρά κομμάτια από μία άλλη γλώσσα γιατί πρακτικά και θεωρητικά είναι

αδύνατο να αποκλείσει κανείς αυτά τα στοιχεία. Αυτό που έχει σημασία είναι αν το σώμα κειμένων είναι σχεδιασμένο να περιλαμβάνει μία ή περισσότερες γλώσσες.

Στα πολύγλωσσα κείμενα διακρίνονται σε παράλληλα και συγκρίσιμα σώματα κειμένων. Τα παράλληλα σώματα αποτελούνται από τα ίδια κείμενα σε δύο ή περισσότερες γλώσσες. Έρχεται πρώτο το πρωτότυπο κείμενο και ακολουθούν οι μεταφράσεις του σε άλλες γλώσσες. Βασικό χαρακτηριστικό αυτών των κειμένων είναι η ευθυγράμμιση τους, δηλαδή η μετάφραση του πρωτότυπου κειμένου στις άλλες γλώσσες να γίνεται κατά παραγράφους, προτάσεις και λέξεις έτσι ώστε να γίνεται άμεση σύγκριση των αντίστοιχων γλωσσών.

Τα συγκρίσιμα σώματα κειμένων αναφέρονται σε κείμενα δύο ή περισσότερων γλωσσών που έχουν την ίδια γλωσσική ποικιλία και τις ίδιες προδιαγραφές. Έχουν την ίδια αναλογία από τα ίδια είδη κειμένων στην ίδια περίοδο που λήφθηκε το δείγμα χωρίς η αναφορά να γίνεται αναγκαστικά στο ίδιο περιεχόμενο. Στα συγκρίσιμα σώματα δεν υπάρχει αντιστοίχιση, παρά μόνο γίνεται η σύγκριση ανάμεσα στις γλώσσες με βάση τις επιλογές των λέξεων ή τις δομές της σύνταξης, χωρίς να υπάρχει ο παράγοντας της μετάφρασης. Με αυτό τον τρόπο, δεν υπάρχει η εμπλοκή των δομών ή του λεξιλογίου, τα οποία υπάρχουν αποκλειστικά στη διάλεκτο της μετάφρασης.

1.1.3 Χρονική κάλυψη

Ανάλογα με τη χρονική περίοδο που καλύπτουν τα σώματα κειμένων τα διακρίνουμε σε συγχρονικά, τα οποία είναι κείμενα που αναφέρονται σε μία συγκεκριμένη περίοδο χρόνου, και σε διαχρονικά, κείμενα τα οποία έχουν την ίδια γλώσσα σε διαφορετικές όμως περιόδους χρόνου. Για να μελετηθεί διαχρονικά μία γλώσσα χρειάζονται στοιχεία τα οποία δίνονται από τα διαχρονικά σώματα, τα οποία καλύπτουν διάφορες ιστορικές περιόδους μιας γλώσσας ή την πλήρη εξέλιξή της. Αυτά τα σώματα κειμένων έχουν δημιουργηθεί για αρκετές γλώσσες και κυρίως, για την αγγλική.

1.1.4 Τροπικότητα

Σε αυτή την κατηγορία, ένα σώμα κειμένου μπορεί να έχει τον χαρακτήρα του προφορικού, γραπτού ή μικτού κειμένου. Τα πιο συνηθισμένα είναι τα γραπτά κείμενα γιατί είναι πιο εύκολα διαθέσιμα σε σχέση με τα προφορικά κείμενα. Είναι δύσκολο να γίνει συγκρότηση των προφορικών σωμάτων κειμένων συγκριτικά με τα γραπτά καθώς βασική προϋπόθεση των προφορικών είναι πρώτα να ηχογραφηθούν ή να καταγραφούν τα στοιχεία και μετά να μεταγραφούν σε γραπτό κείμενο για την ευκολότερη πρόσβασή τους. Βασικό στοιχείο τους είναι η συνοδεία του γραπτού αυτού κειμένου με την αντίστοιχη ηχογράφηση του έτσι ώστε να είναι δυνατό να ανατρέξεις στο πρωτότυπο ηχητικό υλικό σε κάθε δυσκολία. Αυτό χρειάζεται, όχι για τον έλεγχο και την επανάληψη, αλλά για να είναι δυνατό να εντοπισθούν στοιχεία τα οποία είναι δύσκολο αποδοθούν σε γραπτή μορφή.

1.1.5 Επισημείωση

Όταν συλλέγονται γλωσσικά δεδομένα χωρίς να τα επεξεργαζόμαστε τότε έχουμε ένα απλό σώμα κειμένων. Ωστόσο, μια συνηθισμένη επεξεργασία των σωμάτων κειμένων μας δίνει τη διαδικασία της επισημείωσης που σημαίνει ότι σε ένα σώμα κειμένων προστίθενται ερμηνευτικές γλωσσικές πληροφορίες. Στο χαρακτηρισμό των σωμάτων κειμένων η επισημείωση θεωρείται μία υποπερίπτωση. Ο χαρακτηρισμός αυτός καθιστά αναγκαίες τις πληροφορίες για να επεξεργασθούν και να μορφοποιηθούν τα κείμενα, για να κατηγοριοποιηθούν σε κειμενικά είδη ή για να συσταθούν τα μέρη τους με τη βοήθεια κάποιας κωδικοποίησης.

1.1.6 Ανανεωσιμότητα

Βασικό στοιχείο των σωμάτων κειμένων είναι ότι έχουν συγκεκριμένα όρια. Τα περισσότερα από αυτά είναι συγκεκριμένα, ολοκληρώνονται σε μια συγκεκριμένη χρονική περίοδο και παραμένουν σταθερά όταν παρέλθει αυτή η στιγμή. Αυτά είναι τα κλειστά σώματα, τα οποία έχουν συγκεκριμένο μέγεθος και είναι σταθερά.

Υπάρχουν όμως και τα ανοιχτά σώματα κειμένων, τα οποία συνεχώς αυξάνονται καθώς προστίθενται σε αυτά καινούρια δεδομένα. Αυτά τα σώματα δίνουν τη δυνατότητα να γίνεται έλεγχος στη μορφή μιας γλώσσας, δίνοντας τη δυνατότητα να προσθέτονται νέα δεδομένα, χωρίς όμως να επηρεάζεται η σύσταση των κειμένων τους. Καθώς ανανεώνονται με νέα κείμενα, μπορεί να αντικαθίστανται τελείως τα παλιά, η σύσταση και το μέγεθος τους όμως παραμένουν σταθερά τοποθετώντας τα παλιά κείμενα σε ένα αρχείο. Μπορεί όμως να προστίθενται επιπλέον τα καινούρια στα παλιά κείμενα και να αποτελούν ένα τεράστιο σώμα κειμένων. Με αυτό τον τρόπο έχοντας μια σταθερή βάση έχουμε τη δυνατότητα να παρακολουθούμε τη διαχρονική εξέλιξη μιας γλώσσας.

1.2 Σχεδιασμός σωμάτων κειμένων

Η αφετηρία για να σχεδιαστούν και να δημιουργηθούν τα διάφορα σώματα κειμένων είναι τα χαρακτηριστικά τους, τα οποία αναλύονται παρακάτω.

1.2.1 Περιεχόμενο

Για να καθορίσουμε το περιεχόμενο ενός σώματος κειμένου θα πρέπει το περιεχόμενο αυτό να επιλεγεί με βάση την επικοινωνία στη γλώσσα της κοινότητας. Όλα αυτά αφορούν την ύπαρξη εξωτερικών κριτηρίων, τα οποία θα πρέπει να είναι λίγα σε αριθμό, διακριτά μεταξύ τους και επειδή μέσα από όλη αυτή τη διαδικασία επιδιώκεται να γίνει η μελέτη της γλωσσικής ποικιλίας, θα πρέπει τα κριτήρια αυτά να την εκφράζουν αποτελεσματικά. Στο σώμα των κειμένων θα πρέπει να υπάρχει ομοιογένεια και θα πρέπει να αποφεύγονται κείμενα που θεωρούνται προβληματικά. Όσο αφορά το σώμα κειμένων αναφοράς πρέπει να γίνονται σαφή τα κριτήρια των κειμένων, τα οποία να συνάδουν με το είδος, το αντικείμενο, το θέμα των κειμένων κλπ.

1.2.2 Μέγεθος

Το ιδανικό μέγεθος ενός σώματος κειμένων δεν είναι εύκολο να καθοριστεί εξ αρχής. Τη δεκαετία του 1970 ο επιθυμητός στόχος σε απόλυτους αριθμούς για το μέγεθος των σωμάτων κειμένων ήταν στο ένα εκατομμύριο λέξεις περίπου για την αγγλική γλώσσα. Αυτός ο στόχος έφτασε στο ένα δισεκατομμύριο λέξεις σήμερα, δηλαδή αυξήθηκε πάνω από χίλιες φορές μέσα σε σαράντα χρόνια. Ενώ ο στόχος των εκατό εκατομμυρίων λέξεων δεν φαίνεται να είναι ανέφικτος για μια γλώσσα όπως η ελληνική.

Σύμφωνα με το υλικό που διατίθεται και με το ερευνητικό πρόγραμμα, θα πρέπει εξ αρχής να καθορίζεται ο στόχος για το ποιο θα είναι το μέγεθος του σώματος κειμένων. Ανάλογα με τα ερωτήματα της έρευνας τα οποία τίθενται, αυτός ο στόχος διαφέρει ανάμεσα σε σώματα κειμένων, εξειδικευμένα ή γενικά.

1.2.3 Τήρηση μεταδεδομένων

Για να συσταθεί ένα σώμα κειμένου και να γίνει ο σχεδιασμός του, θα πρέπει να υπάρχει πρότερη συσχέτιση με το περιεχόμενο και τα στοιχεία του. Η αιτιολογία των παραγόμενων σχετίζεται με τα ήδη υπάρχοντα δεδομένα, επομένως αναφερόμαστε στα λεγόμενα μεταδεδομένα, τα οποία είναι τα δεδομένα που δημιουργούνται για τα υπάρχοντα δεδομένα.

α) Τα εκδοτικά μεταδεδομένα είναι τα δεδομένα που σχετίζονται με την αρχική τους πηγή.

β) Τα αναλυτικά μεταδεδομένα είναι αυτά που αναλύουν τα συστατικά μέρη των σωμάτων κειμένων.

γ) Τα περιγραφικά μεταδεδομένα περιγράφουν εσωτερικά ή εξωτερικά χαρακτηριστικά από τα συστατικά μέρη των σωμάτων κειμένων.

δ) Τα διοικητικά μεταδεδομένα είναι τα δεδομένα που ασχολούνται αποκλειστικά με τα σώματα κειμένων.

1.2.4 Δειγματοληψία

Η δειγματοληψία συνδέεται άμεσα με την αντιπροσωπευτικότητα. Για να μπορέσει να καλυφτεί αποτελεσματικότερα η δειγματοληψία, θα πρέπει να επιλεγθούν μικρά δείγματα από πολλά κείμενα και όχι ολόκληρα σώματα κειμένων από λίγα κείμενα. Με τη δειγματοληψία όμως υπάρχει κίνδυνος, όσο προσεκτικά κι αν γίνει αυτή η επιλογή δειγμάτων από κείμενα, να διαστρεβλώσει την εικόνα του κειμένου και συνεπώς της γλώσσας, γιατί μπορεί αυτή η επιλογή να μη συμπεριλάβει τον πρόλογο ή τον επίλογο ενός γραπτού κειμένου.

1.2.5 Αντιπροσωπευτικότητα

Μέσω της αντιπροσωπευτικότητας προσφέρεται μια αντιπροσωπευτική εικόνα της γλωσσικής ποικιλίας που συλλέγεται μέσα σε κάθε σώμα κειμένων. Το θέμα της αντιπροσωπευτικότητας είναι πολύπλοκο, καθώς η έννοια *αντιπροσωπευτικό δείγμα* είναι ένα ζήτημα που απασχολεί τους ερευνητές, οι οποίοι έχουν τοποθετηθεί με διαφορετικό τρόπο σε αυτό. Η έννοια της αντιπροσωπευτικότητας δεν μπορεί να διατυπωθεί με σαφήνεια, διότι θα πρέπει να γνωρίζουμε εκ των προτέρων πόσα και ποια κείμενα θα πρέπει να συλλέξουμε για να πετύχουμε το στόχο μας.

1.2.6 Διαθεσιμότητα

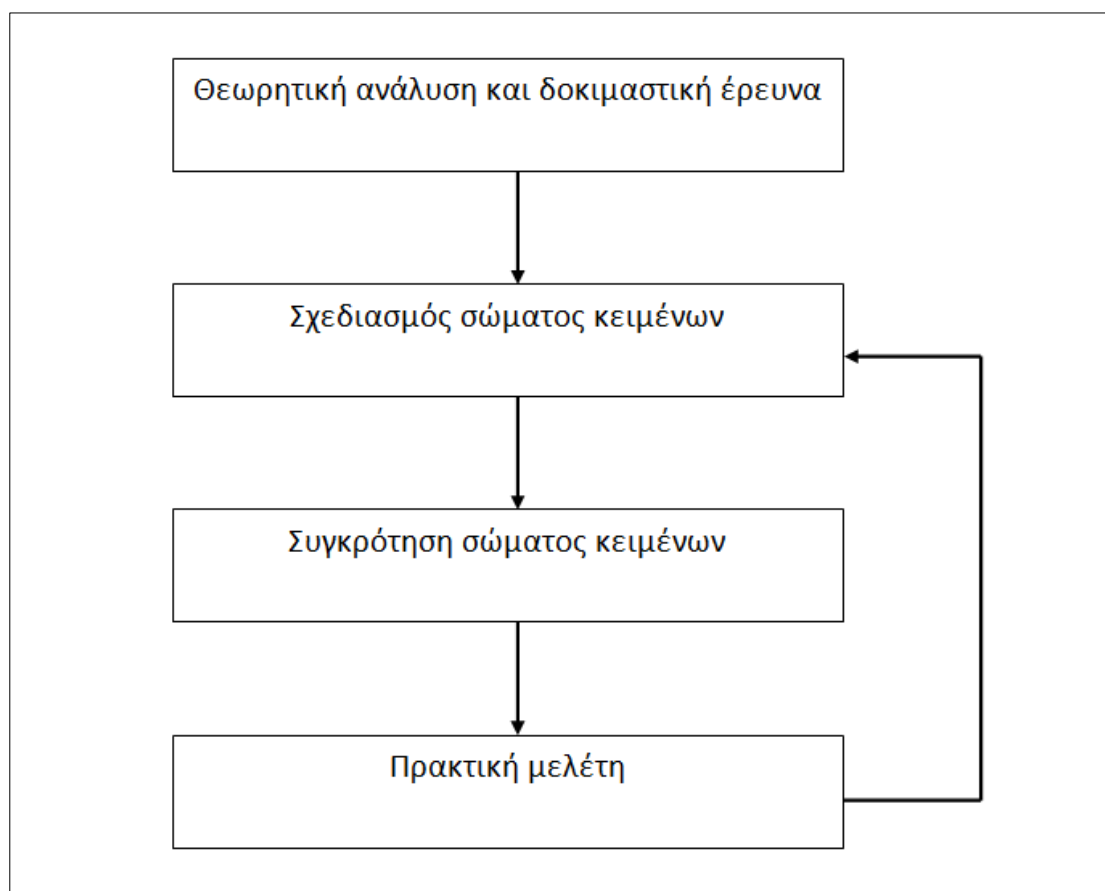
Τα σώματα κειμένων σχετίζονται με τη διαθεσιμότητα του γλωσσικού κειμένου. Αυτό θεωρείται ως ένα πλεονέκτημα, επειδή επιτρέπει να γίνεται έλεγχος των ερευνητικών υποθέσεων καθώς και να επαναλαμβάνονται. Έτσι, μπορούμε να βρίσκουμε αυτά τα σώματα κειμένων μέσω του διαδικτύου, με μόνο περιορισμό τα πνευματικά δικαιώματά τους. Λόγω αυτών, λοιπόν, είναι απαραίτητη η έγγραφη άδεια για να χρησιμοποιηθούν αυτά τα στοιχεία από τους ιδιοκτήτες αυτών των κειμένων. Μια άλλη παράμετρος είναι η δεοντολογία ότι πρέπει να τηρούνται τα προσωπικά δεδομένα των δημιουργών ή εκδοτών. Σε αυτή την περίπτωση, θα πρέπει να έχουμε τη συγκατάθεση των ατόμων που συμμετέχουν σε αυτά τα στοιχεία ή σε προφορικές συνομιλίες, μέσω ενός τρόπου ώστε να τηρούνται η ανωνυμία και τα προσωπικά στοιχεία τους.

Όσο αφορά τα πνευματικά δικαιώματα, αυτά εξαρτώνται από τους νόμους της κάθε χώρας και μπορούμε να τα παρακάμψουμε αυτά, σε περίπτωση που τα γλωσσικά στοιχεία δεν είναι αυτούσια αλλά παραλλαγή αυτών των στοιχείων. Πολλές φορές χρησιμοποιούνται

μερικά κομμάτια κειμένων σε κάποια σώματα, έτσι ώστε να μην μπορούν να χρησιμοποιηθούν με διάφορους τρόπους εκμετάλλευσης. Όταν μέσω του διαδικτύου κυκλοφορεί γλωσσικό υλικό, το θέμα των πνευματικών δικαιωμάτων είναι πολύ μπερδεμένο. Όταν όμως χρησιμοποιούνται σώματα κειμένων για να γίνουν διάφορες έρευνες, η χρήση αυτή θεωρείται νόμιμη αλλά με όρους πάλι.

1.3 Συγκρότηση σωμάτων κειμένων

Για να σχεδιάσουμε ένα σώμα κειμένων δεν είναι απαραίτητο να είναι από πριν διαμορφωμένα τα κριτήρια. Κάθε σώμα κειμένων βασίζεται σε γενικές αρχές που είναι βασισμένες στις ιδιαίτερες συνθήκες της δημιουργίας του. Επομένως, τα στάδια σχεδιασμού και συγκρότησης σωμάτων κειμένων δεν μπορούν να διακριθούν απόλυτα το ένα από το άλλο. Για αυτό τον λόγο, η σχέση τους μπορεί να αναπαρασταθεί στο Σχήμα 1.1 ως κυκλική.



Σχήμα 1: Η ανατροφοδοτούμενη σχέση του σχεδιασμού και της συγκρότησης σωμάτων κειμένων.

Όπως παρατηρούμε στο παραπάνω σχήμα, για τη δημιουργία του σώματος κειμένων αρχικό στάδιο είναι η δοκιμαστική έρευνα και η θεωρητική ανάλυση κάποιων παραμέτρων. Στην πορεία, ακολουθεί ο σχεδιασμός του σώματος κειμένων, βάση του οποίου συγκροτείται ένα μέρος του. Τέλος, η πρακτική μελέτη που συνεπάγεται τη συγκρότηση του σώματος κειμένων οδηγεί στον ανασχεδιασμό αυτού. Επομένως, γίνεται ανατροφοδότηση της όλης διαδικασίας μέχρις ότου προσαρμοστεί στις κατάλληλες συνθήκες.

Αναλυτικότερα, στη συγκρότηση του σώματος κειμένων διερευνώνται πηγές και αναπτύσσονται οι γλωσσικοί πόροι που συνδέονται με το σώμα. Έπειτα, εισάγονται δεδομένα στο σύστημα με τρεις τρόπους: προσαρμόζονται δεδομένα σε ηλεκτρονική μορφή, γίνεται οπτική σάρωση και δακτυλογράφηση για γραπτά δεδομένα, ενώ είναι απαραίτητη και η βιντεοσκόπηση ή μαγνητοφώνηση για προφορικά δεδομένα. Δεδομένα σε ηλεκτρονική μορφή είναι η ευκολότερη πηγή δεδομένων. Γενικότερα όμως, ένας καλός σαρωτής και ένα πρόγραμμα οπτικής αναγνώρισης χαρακτήρων μπορούν να συμβάλλουν στη μετατροπή δεδομένων με τη βοήθεια της οπτικής σάρωσης.

Συνειδητά λοιπόν, η εισαγωγή των δεδομένων με πληκτρολόγηση θεωρείται η πιο χρονοβόρα διαδικασία και έχει το μεγαλύτερο κόστος συγκριτικά με άλλες μεθόδους. Υπάρχουν όμως περιπτώσεις που την καταστούν αναγκαία, όταν η εκτύπωση του πρωτότυπου κειμένου έχει χαμηλή ποιότητα ή όταν κάποια γραπτά κείμενα είναι χειρόγραφα.

Επιπλέον, στη μεταγραφή προφορικών δεδομένων οφείλουμε να λάβουμε υπόψη κάποιες συμβάσεις για την επικάλυψη ομιλητών. Δεν είναι απαραίτητες οι λεπτομερείς πληροφορίες για προσωδιακά φαινόμενα. Παράλληλα, είναι καλό όλα τα συστήματα μεταγραφής να έχουν κάποιες γενικές αρχές, με αποτέλεσμα να επιτυγχάνεται η σύνδεση των διαφόρων προφορικών σωμάτων κειμένων.

Η επόμενη διαδικασία στη συγκρότηση σωμάτων κειμένων είναι ο καθαρισμός των σωμάτων κειμένων από περιττά στοιχεία, από στοιχεία που δεν είναι λεκτικά, όπως εικόνες, κόμματα, κενά κλπ. Γενικότερα, αφορά κάθε είδους διαδικασία μορφοποίησης του αρχικού κειμένου ή κωδικοποίησης μιας ιστοσελίδας σε HTML. Η μετατροπή από την υπάρχουσα μορφοποίηση καθίσταται αναγκαία, καθώς τα λογισμικά επεξεργάζονται κείμενα που είναι αποθηκευμένα σε μορφή txt με κωδικοποίηση UTF-8.

Ακολουθεί η διαδικασία της τυποποίησης του σώματος κειμένων. Τα δομικά μέρη του κειμένου έχουν κάποια βασική τυποποίηση, καθώς και ορθογραφική κανονικοποίηση για τα προφορικά σώματα κειμένων.

Συμπληρωματικά, υπάρχει η διαδικασία της κωδικοποίησης κατά την οποία κωδικοποιούνται και αποθηκεύονται τα στοιχεία της ταυτότητας του κειμένου, στην κεφαλίδα του κειμένου ή σε ανεξάρτητη βάση δεδομένων. Έτσι, επιτρέπεται η αναζήτηση των κειμένων αυτών με βάση κάποια κριτήρια. Αυτά τα κριτήρια οφείλουν να είναι ευέλικτα, έτσι ώστε να είναι δυνατή η πολλαπλή και διαφοροποιημένη σύνθεση σωμάτων, σύμφωνα βέβαια με τις ανάγκες της κάθε έρευνας.

Τέλος, ακολουθεί η διαδικασία της γλωσσικής επισημείωσης του σώματος κειμένων. Με αυτό τον τρόπο, το κείμενο έχει “καθαριστεί” από μη γλωσσικές πληροφορίες κι έτσι, γίνεται πιο αντιληπτή η αναγνωσιμότητα του κειμένου.

Ως συμπλήρωμα, καλό θα ήταν να αναφέρουμε ότι πληθώρα διαφορετικών σωμάτων κειμένων έχουν αναπτυχθεί τις τελευταίες πέντε δεκαετίες, από τότε δηλαδή που δημιουργήθηκε το πρώτο ηλεκτρονικό σώμα κειμένων. Στη σημερινή εποχή, τα αγγλικά

σώματα κειμένων έχουν ξεπεράσει τις δύο δισεκατομμύρια λέξεις. Για παράδειγμα, το [Oxford English Corpus](#) έχει πάνω από δυόμισι εκατομμύρια λέξεις και το [Cambridge English Corpus](#) έχει περίπου τρία δισεκατομμύρια λέξεις. Αντιθέτως, τα ελληνικά σώματα κειμένων είναι μικρά σε μέγεθος και δεν έχουν αναπτυχθεί αρκετά μεγάλα στο διαδίκτυο. Επιπλέον, τα περισσότερα σώματα κειμένων που είναι πολύγλωσσα και ταυτόχρονα περιλαμβάνουν και την ελληνική γλώσσα, έχουν σχέση με κείμενα της Ευρωπαϊκής Ένωσης. Παράλληλα με όλα αυτά τα σώματα κειμένων έχουν αναπτυχθεί κάποια μικρότερα σώματα κειμένων, τα οποία είναι εξειδικευμένα για ειδικούς σκοπούς.

ΚΕΦΑΛΑΙΟ 2

ΤΑ ΓΛΩΣΣΙΚΑ ΕΠΙΠΕΔΑ

Στη σημερινή εποχή, η ανθρώπινη γλώσσα έχει καταστεί αναγκαία για την επικοινωνία των ανθρώπων. Αποτελεί το πιο αναπτυγμένο φυσικό σύστημα επικοινωνίας, ένα σύστημα ήχων και εννοιών. Θεωρείται ένας κώδικας του οποίου τα βασικά γνωρίσματα είναι οι φθόγγοι και οι σημασίες. Η δομή χαρακτηρίζεται από τους φθόγγους, ενώ το περιεχόμενο από τις σημασίες.

Η ανθρώπινη γλώσσα επινοήθηκε ως ένας τρόπος έκφρασης και επικοινωνίας των ανθρώπων. Χωρίζεται σε πέντε επίπεδα σύνταξης, όπως αυτά προκύπτουν από τις διαδικασίες ανάλυσής της.

2.1. Φωνολογικό επίπεδο

Σε αυτό το επίπεδο η φωνολογία περιγράφεται ως ένα σύστημα ήχων και ηχητικών ακολουθιών που χαρακτηρίζουν τη γλώσσα.

α) Ο σκοπός της είναι να διακρίνει την κάθε λέξη σύμφωνα με τα διάφορα φωνήματα. Μπορεί να διαχωρίζει τις λέξεις μόνο από ένα γράμμα, σύμφωνο ή φωνήεν, καθώς επιτρέπει επίσης και την ελεύθερη φωνητική σύνταξη κάποιων λέξεων ως προς ορισμένα φωνήματα.

β) Επιτρέπει τη συνεχή ακολουθία συμφώνων με κάποια παρεμβολή φωνήεντος και συγκεκριμένα το συνδυασμό κάποιων ειδικών φθόγγων, π.χ. σπίτι, στοίχος, σχοινί.

γ) Συνδέεται άμεσα με φωνολογικά στοιχεία τα οποία προσδίδουν τον τονισμό μιας γλώσσας. Ο τονισμός αποτελεί βασικό συστατικό της κάθε γλώσσας στην ύπαρξη ομιλίας.

δ) Αναζητά ιδιότητες οι οποίες είναι κοινές σε όλες τις γλώσσες, ιδιότητες που φωνολογικά έχουν κάποιο κοινό σημείο. Για παράδειγμα, το φωνήεν α υπάρχει σε όλες τις γλώσσες.

2.2. Μορφολογικό επίπεδο

Το μορφολογικό επίπεδο ασχολείται με τα μορφήματα και τις λέξεις της γλώσσας. Τα μορφήματα είναι τα μικρότερα συστατικά, είτε γραμματικά είτε λεξικά, τα οποία δίνουν την μορφή για την σύσταση μιας λέξης. Ως κύρια ζητήματα αυτού του επιπέδου θεωρούνται:

α) η αναγνώριση της λέξης: οι ομιλητές μπορούν να αναγνωρίσουν εύκολα και διακριτά την κάθε λέξη, καθώς αποτελεί κεντρική θέση στη γλώσσα. Σε αυτό το επίπεδο όμως, κατά τη διάρκεια της ανάλυσης τίθεται το ζήτημα των ορίων της λέξης. Για παράδειγμα, μία μονάδα

λέξης μπορεί να υφίσταται ως λέξη μόνη της ή περιλαμβάνονται στα όριά της και οι διάφορες κλίσεις της;

β) η αναγνώριση των μορφημάτων: οι ομιλητές μπορούν να αναγνωρίσουν εξίσου εύκολα και τα μορφήματα, τα οποία αποτελούν τις ελάχιστες υποδιαιρέσεις των λέξεων που διαθέτουν κάποια σημασία. Σε κάποιες περιπτώσεις όμως είναι δύσκολη η αναγνώριση τους:

i. στους ρηματικούς τύπους αναγνωρίζεται το μόρφημα το οποίο δηλώνει παρελθόν, παρόν ή μέλλον. Υπάρχει όμως πρόβλημα στις διάφορες μορφές που μπορεί να έχει ανάλογα με το φωνολογικό περιβάλλον στο οποίο βρίσκεται.

ii. σε κάποιες λέξεις αναγνωρίζονται τα παραγωγικά και κλιτικά επιθήματα, αλλά δεν μπορούν να αναγνωριστούν τα λεξικά μορφήματα (το θέμα της λέξης δηλαδή) καθότι δεν είναι ελληνικής καταγωγής και έχουν μεταφερθεί στα ελληνικά με την ξενόγλωσση προφορά τους.

γ) οι μορφολογικές διαδικασίες: οι διαδικασίες αυτές είναι οι τρεις ακόλουθες:

i. η σύνθεση λεξικών μορφημάτων αφορά την διαδικασία κατά την οποία δημιουργούνται λέξεις με την βοήθεια δύο ή περισσότερων μορφημάτων. Μεταξύ αυτών συνήθως υπάρχει ένα φωνήεν, το οποίο βοηθάει στη σύνδεσή τους ώστε να υφίσταται η νέα λέξη. Οι σύνθετες λέξεις χωρίζονται σύμφωνα με τα συνθετικά τους μέρη σε παρατακτικές, προσδιοριστικές, αντικειμενικές και κτητικές. Οι πρώτες αφορούν τις λέξεις των οποίων τα συνθετικά παρατάσσονται στη σειρά και έχουν δύο σημασίες. Στις προσδιοριστικές το πρώτο προσδιορίζει το δεύτερο συνθετικό, ενώ στις αντικειμενικές το ένα συνθετικό είναι το αντικείμενο του άλλου, προσδιορίζει δηλαδή τη σχέση μεταξύ των δύο συνθετικών. Στις τελευταίες τα συνθετικά προσδιορίζουν ένα χαρακτηριστικό κάποιου ή ένα βασικό γνώρισμα, η σύνθεσή τους δηλώνει την τελική σημασία.

ii. η παραγωγή λέξεων αφορά την διαδικασία κατά την οποία δημιουργούνται λέξεις με τον συνδυασμό λεξικών και παραγωγικών μορφημάτων. Τα τελευταία μπορεί να βρίσκονται πριν ή μετά των λεξικών μορφημάτων. Η παραχθείσα λέξη έχει συνήθως την ίδια σημασία με τα μορφήματα από τα οποία παράγεται. Βασική διαφορά της παραγωγής από την σύνθεση είναι ότι στην παραγωγή τα στοιχεία που παράγουν τις λέξεις δεν μπορούν πάντα να υφίστανται ως λέξεις που έχουν κάποια σημασία.

iii. η κλίση των λέξεων αποτελεί μία βασική διαδικασία για να υφίσταται το νόημα των λέξεων. Με αυτό τον τρόπο διευκρινίζεται ο ενικός ή πληθυντικός αριθμός ενός αντικειμένου, ο χρόνος ύπαρξης (παρελθόν, παρόν, μέλλον), καθώς και η παθητική ή ενεργητική φωνή των ρηματικών τύπων.

δ) η σχέση του με τα υπόλοιπα επίπεδα της γλώσσας: οι μορφολογικές διαδικασίες αλληλεπιδρούν με τα υπόλοιπα γλωσσικά επίπεδα. Στη μορφολογία υπάρχουν κάποιοι ρηματικοί τύποι οι οποίοι έχουν το ίδιο γραμματικό μόρφημα αλλά σε διαφορετική μορφή. Αυτό συμβαίνει καθώς το μόρφημα αλλάζει μορφές ανάλογα με το φωνολογικό περιβάλλον στο οποίο βρίσκεται. Συνεπώς, εμφανίζονται διάφορες μορφές ενός μορφήματος οι οποίες καθορίζονται από την φωνολογία. Το ίδιο συμβαίνει και στη σύνταξη καθώς η κλίση των

λέξεων παρέχει κάποιες συντακτικές πληροφορίες στον χρήστη. Κοινώς, η σύνταξη των λέξεων και κατ' επέκταση των προτάσεων, καθορίζει την μορφή που θα πάρουν οι λέξεις ώστε να δώσουν κάποια συντακτική πληροφορία. Στην περίπτωση της σημασιολογίας οι λέξεις προσαρμόζονται ανάλογα με τη σημασία τους. Τα μορφήματα αποκτούν ή μεταβάλλουν την έννοιά τους ανάλογα με το νόημα που πρέπει να δοθεί. Συνεπώς, το σημασιολογικό επίπεδο συνδέεται άμεσα με το μορφολογικό για την επίτευξη του νοήματος.

ε) η τυπολογία των γλωσσών: οι γλώσσες διακρίνονται ανάλογα με την μορφολογική τους δομή.

i. η αναλυτική γλώσσα αφορά κυρίως τη σύνταξη όπου οι λέξεις λειτουργούν ελεύθερα ως αυτόνομα γραμματικά μορφήματα. Εμφανίζονται με τη σειρά των όρων τους για την επίτευξη νοήματος.

ii. η συντακτική γλώσσα χωρίζεται σε κλιτή και συγκολλητική. Η κλιτή εμφανίζει τις λέξεις με συντακτική σειρά έχοντας ως πρότυπο τα κλιτικά επιθήματα. Στη συγκολλητική γλώσσα τα μορφήματα διατηρούν την σημασία τους και μπορούν να υφίστανται μεμονωμένα στην πρόταση, καθώς επίσης μπορούν να εμφανιστούν και σε διάφορους συνδυασμούς και να έχουν διαφορετική σημασία, αλλά παρόμοια και συνδυαστική με την αρχική τους.

2.3. Συντακτικό επίπεδο

Η σύνταξη είναι το σύνολο των κανόνων που διέπει μια γλώσσα. Χρησιμοποιούνται συντακτικοί κανόνες από τους χρήστες μιας γλώσσας με τέτοιο τρόπο ώστε να γίνονται αντιληπτοί από χρήστες της ίδιας γλώσσας. Συνθέτονται προτάσεις και φράσεις με βάση αυτούς τους κανόνες. Ουσιαστικά, ο συνδυασμός των λέξεων σχηματίζουν τις προτάσεις στις ανθρώπινες γλώσσες.

Από τα αρχαία χρόνια μέχρι σήμερα, έχουν διαμορφωθεί πολλοί συντακτικοί κανόνες και διάφορα πρότυπα για το συντακτικό επίπεδο. Τα περισσότερα από αυτά διαμορφώνουν τα άμεσα συστατικά της πρότασης για αρχή, που είναι οι φράσεις και έπειτα, συνδέουν αυτές μεταξύ τους για να συντάξουν την πρόταση.

2.4. Σημασιολογικό επίπεδο

Σε αυτό το επίπεδο αναφέρεται η σημασία των λέξεων και το περιεχόμενο της γλώσσας. Αφορά δηλαδή τις έννοιες των λέξεων και τις σχέσεις με τις οποίες αυτές συνδέονται. Η σημασία των λέξεων είναι αντικειμενική και για αυτό το λόγο υπάρχουν κάποια λογικά συστήματα που χρησιμοποιούν την αναφορά και τις τιμές αληθείας ώστε να γίνεται πιο ευδιάκριτη η σημασία τους.

Η αναφορά των προτάσεων προσδιορίζει τη σημασία των λέξεων, οι οποίες συνδυάζονται μεταξύ τους με λογικούς συνδέσμους (σύζευξη, άρνηση, τα παράγωγά τους). Η αναφορά

της κάθε λέξης έχει μία σημασία. Ο συνδυασμός αυτών των σημασιών επιφέρει την τελική σημασία της πρότασης. Έπειτα, με τις τιμές αληθείας (οι λεγόμενες “Αλήθεια” και “Ψέμα”) ελέγχουμε κατά πόσο ισχύουν οι προτάσεις. Συνεπώς, ο συνδυασμός των σημασιών των λέξεων επιφέρει την σημασία της πρότασης. Αυτή η σύνθεση σημασιών λέγεται συνθετικότητα.

2.5. Πραγματολογικό επίπεδο

Η πραγματολογία ασχολείται με την επίγνωση των χρηστών για την χρήση της γλώσσας στον κατάλληλο χρόνο, με τις κατάλληλες λέξεις και φράσεις και τον κατάλληλο τρόπο και ύφος. Αφορά κυρίως το επίπεδο εκείνο που υφίσταται στο κοινωνικό περιβάλλον. Τέλος, τα όρια ανάμεσα στο σημασιολογικό και στο πραγματολογικό επίπεδο είναι ασαφή και δεν μπορεί να διευκρινιστεί η διαφορά ανάμεσα στα δύο επίπεδα.

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΥΝΤΑΚΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

3.1 Συντακτική ανάλυση

Η συντακτική ανάλυση εξετάζει τη δομή των προτάσεων με σκοπό να διαπιστώσει αν αυτή είναι σωστή ή λανθασμένη με διάφορους τρόπους. Η πιο απλή προσέγγιση είναι η χρήση ενός τύπου ταιριάσματος προτύπου, με αποτέλεσμα να είναι αναγκαίο να υπάρχουν αποθηκευμένα πρότυπα προτάσεων.

Η συντακτική ανάλυση ομαδοποιεί τις λέξεις βάσει των γραμματικών και συντακτικών κανόνων της γλώσσας. Συνήθως εξετάζεται η συντακτική ανάλυση διαφόρων ερμηνειών των λέξεων κι επιλέγονται εκείνες οι λέξεις που ταιριάζουν καλύτερα στη συντακτική, σημασιολογική και πραγματολογική ανάλυση. Οπότε, τα απαιτούμενα εργαλεία της είναι το λεξικό και η γραμματική.

Ωστόσο, λόγω του μεγάλου αριθμού των λέξεων, δεν δύναται να κρατούνται για κάθε μια λέξη όλες οι μορφές της. Επομένως, αποθηκεύεται μόνο η βασική μορφή κάθε λέξης και οι υπόλοιπες μορφές προκύπτουν με κανόνες μορφολογικής ανάλυσης, για παράδειγμα με βάση την μορφολογία κλίσεων, την ετυμολογική μορφολογία και τη σύνθεση λέξεων.

Κάποιες φορές όμως δεν μπορούν να αναγνωριστούν ορισμένες λέξεις. Η αιτία που συμβαίνει αυτό είναι ότι δεν υπάρχουν οι λέξεις στο λεξικό ή δεν πρόκειται για λέξεις που υφίστανται ή τέλος, έχουν καταχωρηθεί λανθασμένα.

Πιο αποτελεσματική μέθοδος θεωρείται η γραμματική ανάλυση μιας πρότασης, καθώς η πρόταση γίνεται μια ιεραρχική δομή που υποδεικνύει τα συστατικά μέρη της. Η γραμματική ανάλυση γίνεται σε συστήματα, τα οποία αποτελούνται από μια γραμματική και ένα συντακτικό αναλυτή που συγκρίνει την πρόταση με τη γραμματική.

3.1.1 Τα βασικά μέρη της πρότασης

Οι κύριοι όροι της πρότασης είναι το υποκείμενο και το ρήμα ή κατηγορημα. Αυτά είναι τα βασικά μέρη της πρότασης και γύρω από αυτά έπειτα υπάρχουν κι άλλοι όροι. Με τα ρήματα δημιουργούνται φράσεις που έχουν ακέραιο νόημα και λέγονται προτάσεις.

Το υποκείμενο είναι το θέμα για το οποίο γίνεται λόγος στην πρόταση. Πιο πρακτικά, είναι η απάντηση στην ερώτηση “ποιος κάνει αυτό που ορίζει το ρήμα”, π.χ. Η Μαρία έχει μπλε μαλλιά (Ερώτηση: ποιος έχει μπλε μαλλιά. Απάντηση: η Μαρία). Ως υποκείμενα λειτουργούν τα ουσιαστικά, τα ουσιαστικοποιημένα επίθετα, οι αντωνυμίες και άλλες λέξεις, όπως και προτάσεις με πρόταξη του άρθρου ή που εισάγονται με το “όποιος”. Το υποκείμενο συνήθως είναι σε πτώση ονομαστική και σπάνια σε πτώση αιτιατική. Παραλείπεται συχνά, όταν εννοείται από τα συμφραζόμενα ή όταν το υποκείμενο είναι η

προσωπική αντωνυμία. Όπως επίσης, δεν χρησιμοποιείται υποκείμενο στα απρόσωπα ρήματα, π.χ. χιονίζει.

Ρήματα λέγονται οι λέξεις που φανερώνουν πως ένα πρόσωπο, ζώο ή πράγμα ενεργεί ή παθαίνει ή βρίσκεται σε μια κατάσταση. Το ρήμα ονομάζεται και “κατηγόρημα” από την λέξη “κατηγορεί”, καθώς δηλώνει την πράξη του υποκειμένου. Αποτελεί ουσιαστικά τη βάση της πρότασης, την ενέργεια πάνω στην οποία δομείται η πρόταση. Το κατηγόρημα μπορεί να εμφανίζεται με έναν ρηματικό τύπο, οπότε ονομάζεται κατηγορηματικό ρήμα. Μπορεί όμως να εμφανίζεται και με έναν τύπο του ρήματος *είμαι* ή συγγενικού αυτού και με ένα ουσιαστικό ή επίθετο, οπότε το ρήμα λέγεται συνδετικό και το ουσιαστικό ή επίθετο ονομάζεται κατηγορούμενο. Συνδετικά είναι τα ρήματα *είμαι, γίνομαι, φαίνομαι, λέγομαι, θεωρούμαι, ονομάζομαι, ανακηρύσσομαι, κληρώνομαι, μένω, ζω, βρίσκομαι, θγαίνω, στέκομαι* κ.α.

Κατηγορούμενα μπορεί να είναι ουσιαστικά ή επίθετα, μέρη του λόγου όταν λειτουργούν ως ουσιαστικά ή επίθετα, ακόμη και εμπρόθετα μέρη του λόγου. Όταν το κατηγορούμενο προσδιορίζει επιρρηματική σχέση, λέγεται επιρρηματικό κατηγορούμενο, συνήθως όταν στην πρόταση υπάρχει ρήμα που δηλώνει κίνηση.

Το κατηγορούμενο είναι το ουσιαστικό ή το επίθετο που προσδιορίζει το υποκείμενο μέσω του συνδετικού ρήματος. Ενώ το αντικείμενο είναι το αντίστοιχο ουσιαστικό ή επίθετο που προσδιορίζει το υποκείμενο, όταν υπάρχει στην πρόταση ρήμα που δεν είναι συνδετικό.

Το ρήμα μπορεί να λείπει σε κάποιες περιπτώσεις, όταν εκφράζονται έντονες προσταγές, ευχές κτλ. Σε ορισμένες περιπτώσεις που τα ρήματα συντάσσονται με δύο αιτιατικές, η μία αιτιατική προσδιορίζει το αντικείμενο και η άλλη προσδιορίζει ένα κατηγορούμενο του αντικειμένου.

A. Ρήμα με υποκείμενο

- Το ρήμα και το υποκείμενο συμφωνούν ως προς το πρόσωπο και τον αριθμό.
- Αν το υποκείμενο είναι περιληπτικό, τότε το ρήμα μπορεί να βρίσκεται σε πληθυντικό αριθμό.
- Ενώ όταν τα υποκείμενα είναι περισσότερα από ένα, το ρήμα είναι σίγουρα σε πληθυντικό αριθμό.
- Κάποιες φορές, όμως, όταν προηγείται το ρήμα από τα υποκείμενα που είναι σε ενικό αριθμό, τότε το ρήμα είναι και αυτό σε ενικό αριθμό.
- Τέλος, το ρήμα είναι στο επικρατέστερο πρόσωπο, όταν τα υποκείμενα είναι δύο ή περισσότερα και βρίσκονται σε διαφορετικά πρόσωπα. Το επικρατέστερο αυτό πρόσωπο είναι το πρώτο σε σχέση με τα άλλα δύο και το δεύτερο σε σχέση με το τρίτο.

B. Κατηγορούμενο με υποκείμενο

- Όταν το κατηγορούμενο είναι ουσιαστικό, ταιριάζει με την πτώση του υποκειμένου, ενώ όταν είναι επίθετο, ταιριάζει με την πτώση, το γένος και τον αριθμό του υποκειμένου.
- Σπάνια όμως το κατηγορούμενο που είναι επίθετο, μπορεί να είναι στον ενικό αριθμό του ουδέτερου γένους, χωρίς να συμφωνεί με τον αριθμό και το γένος του υποκειμένου.
- Όταν στην πρόταση υπάρχουν περισσότερα από ένα υποκείμενα, τότε το κατηγορούμενο είναι σε πληθυντικό αριθμό.
- Όταν τα υποκείμενα είναι έμψυχα και περισσότερα από ένα, τότε το κατηγορούμενο είναι σε πληθυντικό αριθμό και στο γένος των υποκειμένων. Όταν όμως είναι διαφορετικού γένους τα υποκείμενα, τότε το κατηγορούμενο είναι στον πληθυντικό αριθμό και στο επικρατέστερο γένος. Το επικρατέστερο αυτό γένος είναι το αρσενικό σε σχέση με τα άλλα δύο και το θηλυκό σε σχέση με το ουδέτερο.
- Όταν τα υποκείμενα είναι περισσότερα από ένα και δηλώνουν πράγματα, το κατηγορούμενο δεν εξαρτάται από το γένος των υποκειμένων και είναι πάντα στο ουδέτερο γένος.

3.1.2 Ονοματική και ρηματική φράση

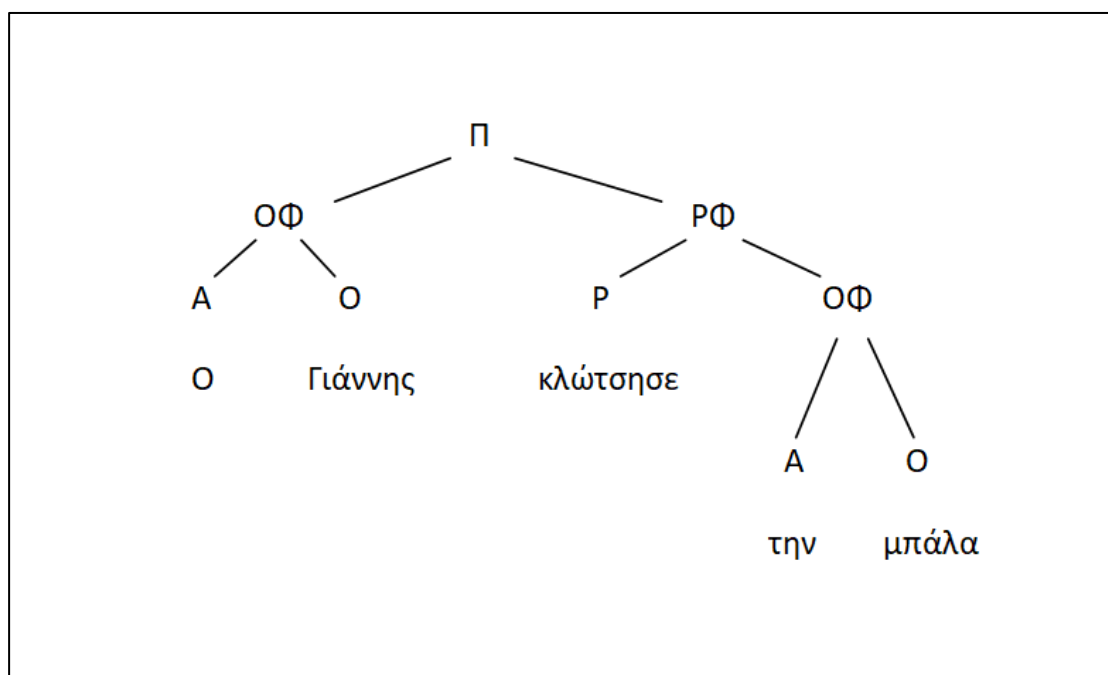
Η ονοματική φράση ορίζεται ως ένα σύνολο λέξεων που μπορούν να αντικατασταθούν από άλλες λέξεις με ισότιμο ρόλο σύνταξης στην πρόταση. Ως πυρήνα έχει ένα ουσιαστικό ή κάθε λέξη που μπορεί να υφίσταται σε θέση ουσιαστικού. Μια ονοματική φράση μπορεί να αποτελείται μόνο από ένα ουσιαστικό, μπορεί όμως να αποτελείται και από ένα ουσιαστικό συνδυασμένο με άλλα μέρη του λόγου. Αυτά τα μέρη του λόγου που συνδυάζεται το ουσιαστικό είναι συνήθως άρθρα, επίθετα, μετοχές, αντωνυμίες, άλλα ουσιαστικά και αριθμητικά. Η ονοματική φράση λειτουργεί ως υποκείμενο, αντικείμενο ή κατηγορούμενο, προσδιορισμός άλλης ονοματικής φράσης ή ρήματος.

Το ρήμα θεωρείται το βασικό συστατικό της πρότασης, πάνω στο οποίο στηρίζεται το νόημα αυτής. Το ρήμα εκφράζει την ενέργεια ή την κατάσταση στην οποία βρίσκεται το υποκείμενο της πρότασης. Τέλος,, το ρήμα καθορίζει τον αριθμό και τα είδη των γλωσσικών στοιχείων που ακολουθούν.

Η ρηματική φράση ορίζεται ως το σύνολο των λέξεων που αποτελείται από το ρήμα και τα στοιχεία που το συμπληρώνουν, δηλαδή το αντικείμενο ή κατηγορούμενο και τα επιρρηματικά. Μπορεί όμως να αποτελείται μόνο από το ρήμα. Επίσης, υπάρχουν περιπτώσεις που τα λεκτικά στοιχεία μιας ρηματικής φράσης μπορούν να ενωθούν και να αποδοθούν με ένα μόνο ρήμα. Τέλος, η ρηματική φράση έχει τον ρόλο της επεξήγησης και της ολοκλήρωσης του νοήματος μιας πρότασης.

Το υποκείμενο και το αντικείμενο είναι δύο λέξεις που αποτελούνται από παρόμοια χαρακτηριστικά. Αυτά τα χαρακτηριστικά είναι τα άρθρα, τα ουσιαστικά κτλ.

Για παράδειγμα:



Σχήμα 2: Παράδειγμα με την πρόταση “Ο Γιάννης κλώτσησε την μπάλα”.

Τα λεξικά στοιχεία ο και την ανήκουν στην ίδια κατηγορία, την κατηγορία του άρθρου. Οι λέξεις Γιάννης και μπάλα ανήκουν επίσης στην ίδια κατηγορία ως ονόματα (κύριο και κοινό), στην κατηγορία του ουσιαστικού. Επιπλέον, ο συνδυασμός ενός άρθρου και ενός ουσιαστικού θεωρείται μια φράση, η οποία ονομάζεται ονοματική φράση. Ο συνδυασμός ενός ρήματος με μία ονοματική φράση, όπως φαίνεται στο σχήμα 3.1, θεωρείται μια ρηματική φράση.

Το ρήμα, το άρθρο και το ουσιαστικό είναι οι κατηγοριοποιήσεις που εκφράζουν συντακτικά χαρακτηριστικά, για αυτό το λόγο ονομάζονται συντακτικές κατηγορίες. Επομένως, αν αλλάξουμε τις λέξεις και διατηρήσουμε τις ίδιες συντακτικές κατηγορίες, θα έχουμε άλλο περιεχόμενο πάνω στην ίδια δομή, δηλαδή άλλη πρόταση και συνεπώς, άλλο νόημα.

Συμπερασματικά, ως φυσικοί ομιλητές κατηγοριοποιούμε τις λέξεις που χρησιμοποιούμε και δημιουργούμε νέες φράσεις πάνω στις ίδιες δομές. Η κατηγοριοποίηση αυτή φανερώνει και την παρόμοια συμπεριφορά που έχουν τα στοιχεία μιας κατηγορίας.

3.1.3 Συντακτικές κατηγορίες

Α. Υλικά σύνταξης

Η ελληνική γλώσσα έχει δέκα είδη λέξεων, που ονομάζονται μέρη του λόγου. Σε αυτά ανήκουν όλες οι λέξεις της γλώσσας και χωρίζονται σε κλιτά και άκλιτα μέρη του λόγου. Τα μέρη του λόγου που κλίνονται, δηλαδή αλλάζουν μορφή μέσα στο λόγο, είναι κλιτά. Αυτά που δεν κλίνονται, δεν αλλάζουν μορφή μέσα στο λόγο, λέγονται άκλιτα μέρη του λόγου. Τα κλιτά είναι το άρθρο, το ουσιαστικό, το επίθετο, η αντωνυμία, το ρήμα και η μετοχή, ενώ τα άκλιτα μέρη είναι το επίρρημα, η πρόθεση, ο σύνδεσμος και το επιφώνημα. Το ουσιαστικό και το επίθετο λέγονται και ονόματα. Η μετοχή όμως δεν είναι μια ξεκάθαρη περίπτωση, καθώς οι σχηματισμοί *-μένος, -η, -ο* είναι κλιτές μετοχές και λειτουργούν ως επίθετα, ενώ οι σχηματισμοί *-οντας, -ώντας* είναι άκλιτες μετοχές και λειτουργούν ως επιρρήματα.

Οι μορφές, λοιπόν, που μπορεί να πάρει μια κλιτή λέξη μέσα στο λόγο λέγονται τύποι της λέξης αυτής. Οι τύποι της λέξης καθορίζονται από την κατάληξη, το θέμα και τον χαρακτήρα. Κατάληξη λέγεται το τελευταίο μέρος μιας κλιτής λέξης, αυτό που αλλάζει. Το θέμα είναι το πρώτο μέρος μιας κλιτής λέξης, αυτό που είναι στην αρχή και δεν αλλάζει. Τέλος, ο χαρακτήρας είναι ο τελευταίος φθόγγος του θέματος μιας λέξης.

Κάθε κλιτή λέξη, εκτός από το ρήμα, παρουσιάζει διάφορους τύπους, που λέγονται πτώσεις. Ουσιαστικά, το άρθρο, το επίθετο, η αντωνυμία, η μετοχή και το ουσιαστικό έχουν πτώσεις και για αυτό το λόγο, λέγονται πτωτικά. Οι πτώσεις είναι τέσσερις:

- Ονομαστική: η πτώση με την οποία απαντούμε στην ερώτηση *Ποιος; Τι;*
- Γενική: η πτώση με την οποία απαντούμε στην ερώτηση *Τίνος;*
- Αιτιατική: η πτώση με την οποία απαντούμε στην ερώτηση *Ποιον; Τι;*
- Κλιτική: η πτώση με την οποία καλούμε κάποιον.

Τα πτωτικά έχουν γένος, αριθμό και κλίση. Τα γένη των πτωτικών είναι τρία:

- Αρσενικό: το γένος που τα πτωτικά μπορούν να πάρουν το άρθρο *ο*.
- Θηλυκό: το γένος που τα πτωτικά μπορούν να πάρουν το άρθρο *η*.
- Ουδέτερο: το γένος που τα πτωτικά μπορούν να πάρουν το άρθρο *το*.

Τα πτωτικά έχουν δύο αριθμούς, τον ενικό και τον πληθυντικό και για κάθε αριθμό έχουν ξεχωριστούς τύπους. Οι τύποι των πτωτικών που δηλώνουν το ένα είναι ενικού αριθμού, ενώ αυτοί που δηλώνουν τα πολλά είναι πληθυντικού αριθμού.

Τα πτωτικά, ανάλογα με τον τρόπο που σχηματίζουν τις πτώσεις τους, χωρίζονται σε κατηγορίες, οι οποίες λέγονται κλίσεις.

Επιπρόσθετα, οφείλουμε να επισημάνουμε ότι οι παραπάνω συντακτικές κατηγορίες διαφοροποιούνται και ως προς το είδος της κλίσης και ως προς τη σημασιολογία τους. Σαφώς, το ρήμα και το ουσιαστικό που είναι και τα δύο κλιτά, δεν έχουν την ίδια κλίση. Επίσης, το ουσιαστικό και το άρθρο που είναι και τα δύο κλιτά, δεν μπορούν να έχουν την ίδια λειτουργία. Το πρώτο μπορεί να φτιάξει από μόνο του μια ονοματική φράση, ενώ το άρθρο δεν μπορεί, παρά μόνο όταν συνδυάζεται μαζί με το ουσιαστικό (π.χ. Το παιδί).

Με βάση τα παραπάνω, διακρίνονται δύο βασικές κατηγορίες, οι λεξικές και οι γραμματικές ή λειτουργικές. Οι πρώτες κατηγορίες είναι πρωταρχικής σημασίας, ενώ οι δεύτερες είναι δευτερεύουσας σημασίας. Αυτό συμβαίνει, επειδή οι λεξικές κατηγορίες ορίζουν τη λέξη της φράσης που δίνει το νόημα ή μια οντότητα (π.χ. ποδήλατο), ενώ οι γραμματικές κατηγορίες έχουν πιο λειτουργικό ρόλο, αποδίδουν δηλαδή αναφορά στη λέξη της φράσης που δίνει το νόημα (π.χ. Το).

Συμπληρωματικά, να αναφέρουμε ότι χαρακτηρίζονται ως ανοιχτή τάξη οι λεξικές κατηγορίες, καθώς έχουμε τη δυνατότητα να δημιουργήσουμε καινούρια ουσιαστικά κατά τη διάρκεια της φυσικής ομιλίας. Αντιθέτως, οι γραμματικές κατηγορίες χαρακτηρίζονται ως κλειστή τάξη, διότι δεν μπορούμε να δημιουργήσουμε νέα άρθρα ή νέα μορφήματα της ίδιας κατηγορίας.

Με λίγα λόγια, οι λεξικές κατηγορίες προσδιορίζονται από τις λειτουργικές, ενώ οι γραμματικές κατηγορίες λειτουργούν πάνω στις λεξικές. Οι λεξικές είναι, συνήθως, τα ουσιαστικά, τα επίθετα, οι προθέσεις, τα ρήματα και τα επιρρήματα. Ως λειτουργικές κατηγορίες αναγνωρίζονται, συνήθως, οι αντωνυμίες, τα άρθρα, οι τύποι της κλίσης, οι δείκτες της άρνησης, οι σύνδεσμοι, καθώς επίσης και τα επιφωνήματα και τα μορφήματα που συνοδεύουν ρηματικούς τύπους (π.χ. θα, ας).

B. Αναγνώριση συστατικών

Όπως αναφέραμε και παραπάνω, μπορούμε να ορίσουμε τις φράσεις σύμφωνα και με τα συστατικά της πρότασης. Αυτή η δυνατότητα υπάρχει, διότι τα λεξικά στοιχεία σχηματίζουν κάποιες ενότητες, δεν μπαίνουν δηλαδή σε τυχαία σειρά. Αυτές οι ενότητες που σχηματίζονται, ονομάζονται συστατικά και δημιουργούν τις φράσεις της πρότασης. Αν συνδυάσουμε, λοιπόν, αυτές τις φράσεις, δημιουργούμε μια πρόταση, η οποία θεωρείται μία μεγάλη φράση.

Ως εκ τούτου, υπάρχουν κάποια κριτήρια αναγνώρισης αυτών των συστατικών, δηλαδή υπάρχει η δυνατότητα να ελέγξουμε και να αποδείξουμε αν μια σειρά λέξεων είναι συστατικό ή όχι. Αυτό όμως δεν είναι εφικτό για όλα τα συστατικά και, σε μεγάλο βαθμό. Αυτά τα κριτήρια είναι το κριτήριο της κατανομής, της αντικατάστασης, της μετατόπισης, της παράλειψης και της σύζευξης.

Ένα κριτήριο αναγνώρισης των συστατικών μπορεί να θεωρηθεί αυτό της κατανομής. Το κριτήριο κατανομής είναι αυτό που παρατηρεί τη θέση της φράσης. Αν μια φράση είναι, για

παράδειγμα, ονοματική, τότε θα πρέπει να εμφανίζεται σε θέσεις όπου βρίσκονται οι ονοματικές φράσεις. Το ίδιο, λοιπόν, ισχύει και για τη ρηματική φράση, την προθετική κλπ.

Ένα άλλο κριτήριο είναι το κριτήριο της αντικατάστασης. Αντικαθίσταται δηλαδή μια φράση από τους αντωνυμικούς τύπους που ανήκουν στην αντίστοιχη κατηγορία με τη φράση. Για παράδειγμα, μια ονοματική φράση μπορεί να αντικατασταθεί από μία αντωνυμία, μία ρηματική φράση μπορεί να αντικατασταθεί από έναν αντίστοιχο τύπο και μια προθετική φράση από ένα επίρρημα τόπου.

Το τρίτο κριτήριο που θα αναφέρουμε είναι το κριτήριο της μετατόπισης. Αυτό σημαίνει ότι αν μια σειρά λέξεων αποτελεί συστατικό, τότε έχει τη δυνατότητα να μετατοπιστεί μέσα στην πρόταση και το αποτέλεσμα αυτό είναι γραμματικό. Δεν μπορεί όμως να διασπαστεί, μετατοπίζεται ολόκληρο το συστατικό όπως είναι, αλλιώς το αποτέλεσμα θεωρείται αντιγραμματικό.

Ένα ακόμη κριτήριο αναγνώρισης συστατικών είναι το κριτήριο της παράλειψης. Όταν έχουμε μια σειρά λεξικών στοιχείων που αποτελούν ένα συστατικό, τότε μπορεί να παραλειφθεί με τον κατάλληλο τρόπο. Για παράδειγμα, όταν έχουμε μια δευτερεύουσα πρόταση που λειτουργεί ως αντικείμενο του ρήματος, τότε μπορεί να θεωρηθεί ως συστατικό. Συνεπώς, μπορεί να παραλειφθεί κιόλας σε περίπτωση που υπάρχει μια δεύτερη πρόταση, στην οποία όμως υπάρχει το κατάλληλο περιβάλλον και το συστατικό ταυτοποιείται με βάση την πρώτη πρόταση.

Ας σημειωθεί, επίσης, το τελευταίο κριτήριο της αναγνώρισης συστατικών, το κριτήριο της σύζευξης. Η σύζευξη συνδέει μόνο όμοια συστατικά, μόνο όμοιες φράσεις. Επομένως, μπορούμε να αντιληφθούμε αν μια σειρά λέξεων είναι συστατικό ή όχι. Σύμφωνα με όλα αυτά, η σύζευξη μπορεί να συνδέσει ονοματικές φράσεις, ρηματικές φράσεις, προθετικές φράσεις, ακόμη και ολόκληρες προτάσεις. Το αποτέλεσμα της σύζευξης όμοιων φράσεων είναι γραμματικό. Σε περίπτωση, λοιπόν, που δεν έχουμε όμοιες φράσεις, μπορούμε αντί για τη σύζευξη να χρησιμοποιήσουμε τη διάζευξη ή την αντίθεση.

Συμπερασματικά, οφείλουμε να τονίσουμε ότι τα κριτήρια αυτά δεν καλύπτουν πάντα όλα τα συστατικά και σε όλες τις περιπτώσεις. Κάποιες φορές, μπορεί τα κριτήρια να μην έχουν ακριβή αποτελέσματα. Κατά συνέπεια, διαπιστώνουμε ότι πρέπει να προσέχουμε τη δομή μας, προτού βγάλουμε συμπεράσματα για τα συστατικά.

3.2 Αλγόριθμοι ταξινόμησης

Η μηχανική μάθηση αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Μπορεί να διεξαχθεί με δύο διαδικασίες, είτε την επαγωγή είτε τον συλλογισμό. Η πρώτη είναι η διαδικασία κατά την οποία από τον ειδικό μεταβαίνουμε στον γενικό κανόνα, ενώ ο συλλογισμός είναι η διαδικασία κατά την οποία από τον γενικό μεταβαίνουμε στον ειδικό κανόνα, δηλαδή το συμπέρασμα.

Στους αλγορίθμους χρησιμοποιούνται κάποια δεδομένα εκπαίδευσης, τα λεγόμενα “παραδείγματα εκπαίδευσης”. Μέσω των αλγορίθμων γίνεται κάποια ανάλυση αυτών των παραδειγμάτων κι έτσι, εκπαιδεύεται το μοντέλο το οποίο θα βοηθήσει σε μια ενδεχόμενη μελλοντική πρόβλεψη. Αυτό, όμως, που συμβάλλει σε μεγάλο βαθμό στα αποτελέσματα των νέων δεδομένων είναι ο domain expert, ο οποίος αποφασίζει ποια θα είναι τα χαρακτηριστικά των παραδειγμάτων εκπαίδευσης που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Συνεπώς, με αυτά τα χαρακτηριστικά εισόδου μπορούμε να συγκρίνουμε τα νέα παραδείγματα με παλαιότερα και να βγάλουμε συμπεράσματα, δεδομένα εξόδου.

3.2.1 Επιβλεπόμενη μάθηση

Η επιβλεπόμενη μάθηση είναι ένα είδος της μηχανικής μάθησης, κατά το οποίο εισέρχονται στο σύστημα τα επιθυμητά δεδομένα εισόδου και εξέρχονται τα αντίστοιχα δεδομένα εξόδου. Δηλαδή σε αυτή τη μάθηση γνωρίζουμε ήδη την επιθυμητή τιμή εξόδου. Μετά από ανάλυση αυτών των παραδειγμάτων, παράγεται το μοντέλο μέσω του οποίου μπορούμε πλέον να χαρακτηρίσουμε νέα παραδείγματα, νέα δεδομένα.

Η επιβλεπόμενη μάθηση χωρίζεται σε δύο κατηγορίες: στην ταξινόμηση και στην παλινδρόμηση. Η πρώτη είναι η επιβλεπόμενη μάθηση που έχει ως έξοδο ονοματική μεταβλητή, ενώ η παλινδρόμηση είναι η επιβλεπόμενη μάθηση με έξοδο αριθμητική μεταβλητή.

Η έννοια της ακρίβειας αναφέρεται στο πόσα παραδείγματα είναι όντως θετικά από αυτά που έχει πει ο ταξινομητής ότι είναι, ενώ η ανάκληση αναφέρεται στο πόσα είπε ο ταξινομητής ότι είναι θετικά από αυτά που όντως είναι. Έστω I μία πρότυπη πληροφορία, R το σύνολο των σχετικών της κειμένων και A το σύνολο των κειμένων απάντησης από μία στρατηγική ανάκτησης. Επίσης, έστω Ra το σύνολο των κειμένων που είναι κοινά στα σύνολα R και A , τότε:

$$\text{Ακρίβεια} = |Ra| / |A| \quad \text{και} \quad \text{Ανάκληση} = |Ra| / |R|$$

A. Nearest Neighbors k-NN

Ο αλγόριθμος των πλησιέστερων γειτόνων αφορά τη μάθηση που είναι βασισμένη σε στιγμιότυπα. Στον αλγόριθμο αυτό είναι ήδη καθορισμένος ο αριθμός των πλησιέστερων γειτόνων, ο οποίος είναι συνήθως η μονάδα. Επομένως, γίνεται η αναζήτηση του γείτονα που βρίσκεται πιο κοντά στο παράδειγμα που εξετάζουμε. Με τη βοήθεια του τύπου της ευκλείδειας απόστασης ελέγχουμε όλες τις αποστάσεις μεταξύ των ήδη υπαρχόντων παραδειγμάτων και του παραδείγματος που εξετάζουμε. Η μικρότερη απόσταση αποδίδει τον πλησιέστερο γείτονα. Η ευκλείδεια απόσταση υπολογίζεται από την τετραγωνική ρίζα του αθροίσματος των τετραγώνων της κάθε διαφοράς των χαρακτηριστικών μεταξύ των παραδειγμάτων:

$$EA(1-2) = ((X1 - X2)^2 + (Y1 - Y2)^2 + (Z1 - Z2)^2)^{1/2}$$

*Όταν οι τιμές X1 και X2 είναι αλφαριθμητικές, τότε η διαφορά τους είναι ίση με τη μονάδα αν οι τιμές διαφέρουν ή είναι μηδενική αν οι τιμές είναι ίδιες.

Συμπληρωματικά, μπορώ να βγάλω συμπέρασμα για το παράδειγμα ελέγχου και από το majority voting (ψηφοφορία πλειοψηφίας). Αυτός ο έλεγχος γίνεται στον αλγόριθμο περισσότερων από έναν πλησιέστερων γειτόνων, συνήθως στους τρεις πιο κοντινούς γείτονες. Ελέγχω τα αποτελέσματα αυτών των τριών παραδειγμάτων και έπειτα, μέσα από αυτά συμπεραίνω και το αποτέλεσμα του παραδείγματος ελέγχου.

B. Δέντρα απόφασης

Τα επαγωγικά δέντρα απόφασης έχουν διάφορους αλγορίθμους, οι οποίοι λειτουργούν ως η άμεση και αυτόματη απόφαση από ένα σύνολο δεδομένων. Ο στόχος τους συνήθως είναι να βρουν τη βέλτιστη απόφαση από τα δεδομένα που έχουν. Η ελαχιστοποίηση σφάλματος ή η ελαχιστοποίηση του αριθμού των κόμβων είναι εξίσου σημαντικά και μπορεί να θεωρηθούν επίσης ως στόχοι των αλγορίθμων.

Τα δέντρα απόφασης έχουν δύο μεθόδους επίλυσης των προβλημάτων: top-down και bottom-up. Ο πιο συχνά χρησιμοποιούμενος τρόπος επίλυσης προβλημάτων είναι ο πρώτος. Στα δέντρα απόφασης, με τη μέθοδο top-down κάθε κόμβος διαιρεί τα δεδομένα σε μικρότερα σύνολα.

Ο John Ross Quinlan είναι ο εφευρέτης τριών αλγορίθμων για τα δέντρα απόφασης, id3, C4.5 και Cart.

B.1 id3

Ο αλγόριθμος id3 είναι εύκολος στη χρήση, για αυτό δημιουργείται εύκολα το επιθυμητό δέντρο απόφασης. Δεν μπορεί όμως να τρέξει αν το data set έχει αριθμητικές μεταβλητές. Ο αλγόριθμος αυτός προσπαθεί να προσδιορίσει τον πιο καλό τρόπο κάθε φορά για τη δημιουργία ενός κόμβου. Επομένως, πρέπει να υπολογίζουμε κάθε φορά το κέρδος από την

επιλογή ενός συγκεκριμένου τρόπου. Αυτό σημαίνει ότι υπολογίζεται κάθε φορά κάποια σταθερά, που είναι το κέρδος για κάθε παράμετρο για το δέντρο απόφασης και τέλος, επιλέγεται αυτή η παράμετρος που έχει το κέρδος με τη μεγαλύτερη τιμή. Αυτή είναι που θα μπει στη ρίζα του δέντρου απόφασης, στην κορυφή δηλαδή.

Για να υπολογίσουμε το κέρδος της παραμέτρου, μετράμε την εντροπία. Η εντροπία είναι το σύνολο όλων των πιθανών αποτελεσμάτων μιας κατάστασης ή ενός προβλήματος, δηλώνει την αβεβαιότητα του αποτελέσματος αυτού. Ως συμπλήρωμα έρχεται η πληροφορία, η οποία είναι ουσιαστικά το αποτέλεσμα της κατάστασης, είναι δεδομένο πλέον. Υπολογίζεται ως εξής:

$$I_{(bit)} = \sum_{i=1}^n - p_i * \log_2 p_i$$

B.2 J48

Ο J48 είναι μια παραλλαγή του αλγορίθμου id3 που επιτρέπει την χρήση αριθμητικών μεταβλητών. Συνήθως, το δέντρο απόφασης του αλγορίθμου J48 μπορεί να θεωρηθεί κατάλληλος ταξινομητής, καθώς η ταξινόμηση έχει αρκετά μεγάλη ακρίβεια και παρουσιάζει ελάχιστα λάθος αποτελέσματα.

Γ. Bayes learning (στατιστική μάθηση)

Η μάθηση Bayes παρέχει πρακτικούς αλγορίθμους μάθησης, απαιτεί προηγούμενες πιθανότητες. Επίσης, παρέχει ένα χρήσιμο εννοιολογικό πλαίσιο.

Το θεώρημα Bayes παρουσιάζεται με τον εξής τύπο:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Όπου:

$P(A)$ η προηγούμενη πιθανότητα υποθέσεων A ,

$P(B)$ η προηγούμενη πιθανότητα κατάρτισης δεδομένων B ,

$P(A|B)$ η πιθανότητα να συμβεί το A , δεδομένου ότι έχει συμβεί το B ,

$P(B|A)$ η πιθανότητα να συμβεί το B , δεδομένου ότι έχει συμβεί το A .

Υπάρχει περίπτωση να συμβούν περισσότερα δεδομένα B , οπότε θα υπάρχουν περισσότερες μεταβλητές όπως επίσης και μία κλάση ταξινόμησης C .

Άρα, ο τύπος Bayes γίνεται ως εξής:

$$P(C|a, b, c, \dots) = \frac{P(a, b, c, \dots | C) * P(C)}{P(a, b, c, \dots)}$$

Σε αυτή την περίπτωση αγνοώ τον παρονομαστή και υπολογίζονται οι πιθανότητες $P(a, b, c, \dots | C)$ και $P(C)$ από τα παραδείγματα εκπαίδευσης. Αυτές οι πιθανότητες ονομάζονται παράμετροι του στατιστικού μοντέλου ή στατιστικές παράμετροι.

Επιπρόσθετα, υπάρχει το μοντέλο Naïve Bayes, το οποίο υποστηρίζει την παραδοχή ανεξαρτησίας των δεδομένων, κατά το οποίο η πιθανότητα $P(a, b | C)$ χωρίζεται σε δύο άλλες πιθανότητες, δηλαδή:

$$P(a, b | C) = P(a | C) \times P(b | C) \quad , \quad P(a, b) = P(a) \times P(b)$$

Σε αυτό το σημείο θα αναφερθούμε και σε μία ακόμη παράμετρο, *sparse data problem*, δηλαδή τα προβλήματα σπάνιων φαινομένων. Για όλες τις τιμές των μεταβλητών έχουμε θετικά και αρνητικά αποτελέσματα. Αν τύχει όμως να μην υπάρχει θετικό αποτέλεσμα σε μια τιμή μιας μεταβλητής, τότε η αντίστοιχη πιθανότητα που ορίζει το αποτέλεσμα της ως θετικό είναι μηδενική. Επομένως, έχει καταδικαστεί όλο το αποτέλεσμα και υπολογίζεται ακόμη και η αρχική πιθανότητα ως μηδενική, λόγω της πράξης του πολλαπλασιασμού στον τύπο. Η αντιμετώπιση αυτών των προβλημάτων επιτυγχάνεται με τη μέθοδο της εξομάλυνσης, το *smoothing*.

Για παράδειγμα, στη διαδικασία *add-1 smoothing* γίνεται πρόσθεση μιας μονάδας στον αριθμητή και στον παρονομαστή προστίθεται το πλήθος των τιμών του χαρακτηριστικού της αντίστοιχης πιθανότητας.

3.2.2 Μη επιβλεπόμενη μάθηση

Η μη επιβλεπόμενη μάθηση είναι μία κατηγορία της μηχανικής μάθησης, κατά την οποία γίνεται μια προσπάθεια ανακάλυψης ενός πιθανού μοντέλου. Τα παραδείγματα που χρησιμοποιούνται δεν περιλαμβάνουν την πληροφορία εξόδου, επομένως είναι δύσκολο να αξιολογηθούν οι πιθανές λύσεις του μοντέλου.

Χρησιμοποιείται η διαδικασία της ομαδοποίησης για την μη επιβλεπόμενη μάθηση. Ο αλγόριθμος *k-Means* βασίζεται στον πλησιέστερο γείτονα, όπου κάθε ομάδα ορίζεται από το κέντρο της. Ελέγχω σε κάθε ομάδα για όλα τα σημεία. Ουσιαστικά, χρησιμοποιώ *Manhattan* για να ελέγξω σε ποια κέντρα είναι κοντά, τα χωρίζω σε τρεις ομάδες και μετά, υπολογίζω το κέντρο της κάθε μίας.

$$Manhattan(k-v) = |X1 - X2| + |Y1 - Y2|$$

Με τη συνάρτηση αυτή υπολογίζουμε τις αποστάσεις μεταξύ σημείων.

Τα σημεία του κέντρου ορίζονται από τις τετμημένες και τις τεταγμένες τους. Σε αυτή τη μάθηση δεν υπάρχει ποσοτική αξιολόγηση, αλλά ποιοτική.

ΚΕΦΑΛΑΙΟ 4

ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΤΑΚΤΙΚΩΝ ΣΥΣΧΕΤΙΣΕΩΝ ΣΕ ΑΓΓΛΙΚΑ TWEETS

4.1 Συγκέντρωση αγγλικών tweets

Το πρώτο βήμα της έρευνας ήταν να φτιαχτεί ένας λογαριασμός στο Twitter για να υπάρχει η δυνατότητα πρόσβασης στα δεδομένα. Έπειτα, έγινε ένας λογαριασμός για Twitter Developer, έτσι ώστε να είναι δυνατή η πρόσβαση στο Twitter API. Με αυτό τον τρόπο, δημιουργήθηκε μια εφαρμογή twitter, στην οποία υπάρχουν τα μοναδικά κλειδιά, μυστικά κλειδιά και διακριτικά πρόσβασης. Όλα αυτά χρησιμοποιήθηκαν στον κώδικα παρακάτω για να πιστοποιηθεί η ταυτότητα του χρήστη και να υπάρχει δυνατότητα πρόσβασης στα περισσότερα προϊόντα των προγραμματιστών του Twitter.

Επιπρόσθετα, για να τρέξει ο κώδικας, κατέβηκε η τελευταία έκδοση της python. Με τη βοήθεια αυτού του προγράμματος έτρεξαν τα αρχεία.

Με τον παρακάτω κώδικα, ο οποίος αποθηκεύτηκε σε ένα αρχείο που ονομάζεται streamer, μαζεύτηκαν τα Tweets από το Twitter και τοποθετήθηκαν σε ένα αρχείο που λέγεται out.txt. Χρησιμοποιήθηκε μια Python βιβλιοθήκη που ονομάζεται Tweepy για να γίνει η σύνδεση στο Twitter Streaming API και να κατέβουν τα δεδομένα. Το Tweepy, η βιβλιοθήκη που κατεβάζει τα Tweets βγάζει τα δεδομένα σε μορφή Json.

Επίσης, τα `access_token`, `access_token_secret`, `consumer_key` και `consumer_secret` δημιουργήθηκαν όταν φτιάχτηκε η εφαρμογή Twitter.

Επιπρόσθετα, ένα σημαντικό κομμάτι είναι η εγκατάσταση της εντολής `pip`. Αυτό γίνεται στις ρυθμίσεις της εγκατάστασης της python. Επομένως, υπάρχει η δυνατότητα να ανοιχτεί η γραμμή εντολών του υπολογιστή (`cmd`) και να τρέξει η εντολή `python -m pip install tweepy`. Έτσι, με αυτό τον τρόπο εγκαθίσταται η βιβλιοθήκη `tweepy`, η οποία είναι βασική στον κώδικα για την πρόσβαση στα δεδομένα του twitter.

```

import csv

#kanw installation to tweepy
#pip install tweepy

#kanw import tis vasikes methodous apo to tweepy
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#Edw vazw ta credentials pou pira apo to twitter
access_token = '1058413051379101697-Mr6E5SLRMijlzcOZybu8Y63vh4NYaL'
access_token_secret = 'w3XQhr0OHjCdvNAUw621QnABgRafSPY91xKn3CJSvwdky'
consumer_key = 'u92F8NhCGYjJ254ehFQfrJ0iA'
consumer_secret = 'qBMYqdRymMPyKCmvlSN856ipT4M4Q3L6yT6wCFqBXuRX9w4aPq'

ofp = open('out.csv', 'w', newline='', encoding='utf-8') #edw
orizw to output arxeio. pws 8a onomazetai, se poio proorismo
klp#

ow=csv.writer(ofp)

#auto einai enas vasikos akroaths pou tupwnei mono ta
lamvanomena tweets sto out arxeio
class StdOutListener(StreamListener):
#edw kanw define ton streamer tou twitter
    def on_data(self, data):
        ow.writerow([data])
        return True

    def on_error(self, status):
        print (status)

if __name__ == '__main__':

    #diaxeirizetai tin authentikotita tou Twitter kai tin
    sundesh sto Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    #filtrarei ta tweets gia tin katagrafh dedomenwn apo ta
    agglika
    stream.filter(languages=['en'])

    #mazevei 100 tweets
    num = 100

```

`python streamer.py` και βγαίνουν ως αποτελέσματα τα δεδομένα της μορφής αυτής:

Εικόνα 1: Συγκρότηση σώματος κειμένων από αγγλικά tweets.

35

Με τον παρακάτω κώδικα επιλέγονται τα δεδομένα από το αρχείο out.txt, οι στήλες δηλαδή που είναι αναγκαίες για χρήση στην παρούσα κατάσταση και τοποθετούνται σε ένα νέο αρχείο, το οποίο ονομάστηκε out2.txt.

```
import json

#diavazei apo to json to arxeio
input_file=open('out.json', 'r')

#vgazei auta pou thelw se ena allo arxeio
output_file=open('out2.json', 'w')

#apothikevei to arxiko json arxeio stin mnhmh
json_decode=json.load(input_file)

for item in json_decode:
    my_dict={}
    my_dict['corpus']=item.get('text')

    #ektupwnontai mono ta texts apo ta tweets
    print my_dict
    back_json=json.dumps(my_dict, output_file)

output_file.write(back_json)
output_file.close()
```

4.2 Χαρακτηριστικά μηχανικής μάθησης

Χρησιμοποιούνται όλα αυτά τα tweets που έχουν συλλεχθεί για να φτιαχτεί χειρόγραφα ένα αρχείο excel, στο οποίο θα εκφράζεται η κάθε λέξη που ερευνάται. Χρησιμοποιούνται 1005 παραδείγματα στο αρχείο αυτό, για να γίνουν κάποια πειράματα μηχανικής μάθησης. Με αυτό τον τρόπο εκπαιδεύεται ένας αλγόριθμος ταξινόμησης λέξεων σε ουσιαστικό, αντικείμενο ή σε κανένα από τα δύο. Η εξόρυξη της συντακτικής γνώσης υλοποιείται με τεχνικές της μηχανικής μάθησης.

Η κάθε στήλη του αρχείου excel έχει ένα χαρακτηριστικό, το οποίο παίρνει κάποιες τιμές. Αναφέρεται παρακάτω το κάθε χαρακτηριστικό τι εκφράζει και τι τιμές παίρνει. Το αρχείο φτιάχτηκε χειρόγραφα και τα χαρακτηριστικά υπολογίστηκαν σύμφωνα με την λογική του περιβάλλοντος της λέξης που εξετάζεται. Κοινώς, επιλέχθηκαν τα χαρακτηριστικά που αφορούν την απόσταση των λέξεων γύρω από την κύρια λέξη που ερευνάται, καθώς και τα μέρη του λόγου τους. Με αυτό τον τρόπο θα αναγνωριστούν τα μέρη του λόγου που είναι συνήθως σε θέση υποκειμένου ή αντικειμένου.

Συνεπώς, σημαντικό ρόλο στην αναγνώριση των μερών του λόγου παίζουν το ρήμα και τα μέρη του λόγου γύρω από αυτό. Σύμφωνα με αυτή τη γνώση επιλέχθηκαν τα παρακάτω χαρακτηριστικά για την αναγνώριση συντακτικών συσχετίσεων. Τα χαρακτηριστικά, τα οποία ερευνώνται για τις λέξεις, είναι τα εξής:

A. Η απόσταση από το ρήμα μέχρι τη λέξη που θέλουμε (η αρνητική απόσταση σημαίνει ότι η λέξη μας είναι στην αριστερή πλευρά του ρήματος, τόσες λέξεις πριν το ρήμα όσες και ο αριθμός που λέει). Οι τιμές που παίρνει είναι αριθμητικές.

π.χ. he goes there. Το he είναι στην θέση -1, ενώ το there είναι στην θέση 1.

B. Ο αριθμός άλλων ρημάτων μεταξύ του κύριου ρήματος και της λέξης που θέλουμε. Οι τιμές που παίρνει είναι αριθμητικές.

π.χ. he has been painted. το ρήμα είναι painted. Αν ελέγχουμε το he, τότε υπάρχουν άλλα δυο ρήματα μεταξύ του και του κύριου ρήματος.

C. Ο αριθμός των κομμάτων μεταξύ του ρήματος και της λέξης που θέλουμε. Οι τιμές που παίρνει είναι αριθμητικές.

D. Το ρήμα. Οι τιμές που παίρνει είναι αλφαριθμητικές, δηλαδή λέξεις.

E. Το μέρος του λόγου του ρήματος. Οι τιμές που παίρνει είναι αλφαριθμητικές, δηλαδή συντομογραφίες.

F-I. Οι δύο λέξεις που προηγούνται της λέξης που θέλουμε και τα μέρη του λόγου τους. Οι τιμές που παίρνει είναι αριθμητικές, δηλαδή λέξεις και συντομογραφίες.

J-K. Η ίδια η λέξη που θέλουμε και το μέρος του λόγου της. Οι τιμές που παίρνει είναι αριθμητικές, δηλαδή λέξεις και συντομογραφίες.

L-M. Η λέξη που βρίσκεται δεξιά της λέξης που θέλουμε και το μέρος του λόγου της. Οι τιμές που παίρνει είναι αριθμητικές, δηλαδή λέξεις και συντομογραφίες.

N. Αποτέλεσμα (subject, object, nothing). Οι τιμές που παίρνει είναι αριθμητικές και είναι τρεις: S , O , None.

Τα μέρη του λόγου που αναφέρονται στις στήλες E, G, I, K και M του excel παίρνουν τις τιμές τους από τον παρακάτω πίνακα. Στον πίνακα σημειώνονται αριστερά με συντομογραφία (π.χ. VB, PP, JJ) οι τιμές των μερών του λόγου και δεξιά αναφέρονται οι επεξηγήσεις τους ολογράφως.

Σύμφωνα με τον παρακάτω πίνακα, κατατάχθηκαν οι λέξεις στα μέρη του λόγου τους:

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NP	Proper noun, singular
15.	NPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PP	Personal pronoun
19.	PP\$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VCN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present

Εικόνα 2: Tagset του WSJ Corpus. Πηγή: <https://e-class.ionio.gr/modules/document/document.php?action=view&id=/Weka/POStagging-ask5.pdf>

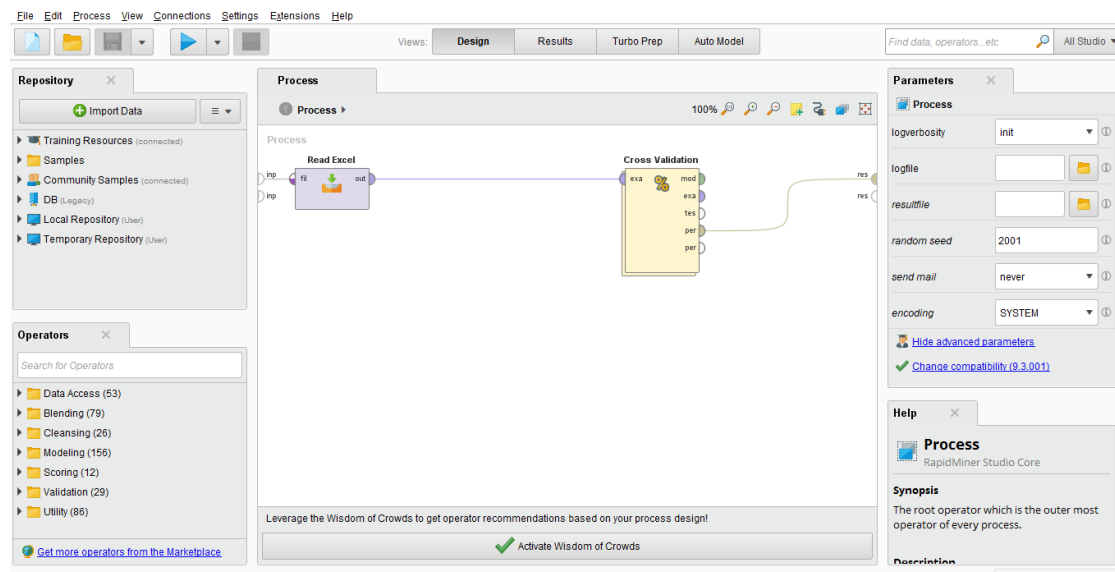
Instances: 1005 παραδείγματα μηχανικής μάθησης

Attributes: 14 χαρακτηριστικά εκπαίδευσης

Όταν το dataset ήταν έτοιμο, το αρχείο excel φορτώθηκε στο Rapidminer για να τρέξουν τα πειράματα ταξινόμησης. Ανοίγοντας την εφαρμογή και ξεκινώντας ένα νέο project, εισάγονται τα δεδομένα με την επιλογή import data. Στη συνέχεια, μέσα στο process επιλέγεται η διαδικασία για να τρέξουν οι αλγόριθμοι με τη βοήθεια του read excel και του cross validation. Στο read excel επιλέγεται το αρχείο excel με τα δεδομένα που είναι προς επεξεργασία και στο cross validation ρυθμίζονται οι αλγόριθμοι με το σωστό performance για να εξαχθούν όλες οι μετρικές των πειραμάτων. Χρησιμοποιήθηκε ως test options το cross validation με 10 folds. Τέλος, με το play τρέχουν οι αλγόριθμοι και δίνονται τα αποτελέσματα.

Για την παραπάνω διαδικασία επιλέχθηκαν οι αλγόριθμοι Naive Bayes, k-πλησιέστεροι γείτονες και Iterative Dichotomiser 3 (ID3). Ο Naive Bayes είναι μια δημοφιλής και βασική μέθοδος για την ταξινόμηση και κατηγοριοποίηση των λέξεων και των κειμένων. Ο αλγόριθμος των k-πλησιέστερων γειτόνων είναι μια μέθοδος που χρησιμοποιείται συχνά για την ταξινόμηση, καθώς η κατηγοριοποίηση των λέξεων γίνεται σύμφωνα με μια πλειονότητα των ψήφων των γειτόνων του, δηλαδή σύμφωνα με την τάξη που είναι πιο συνηθισμένη στους πλησιέστερους γείτονες. Το ID3 είναι ο αλγόριθμος για να δημιουργηθεί ένα δέντρο αποφάσεων από ένα σύνολο δεδομένων και χρησιμοποιείται τυπικά στους τομείς της μηχανικής μάθησης και της επεξεργασίας της φυσικής γλώσσας.

Για την διεκπεραίωση της εργασίας κατέβηκαν στον υπολογιστή τρεις χιλιάδες tweets. Από αυτά όμως χρησιμοποιήθηκαν στην έρευνα μόνο τα εξακόσια tweets, τα οποία είχαν ρήμα στην πρόταση τους. Από αυτά τα tweets φτιάχτηκαν χειρωνακτικά σε ένα excel χίλια πέντε παραδείγματα μηχανικής μάθησης. Τα παραδείγματα αυτά έχουν ως αποτέλεσμα υποκείμενο (S), αντικείμενο (O) και τίποτα από τα δύο (None). Από αυτά τα παραδείγματα μηχανικής μάθησης τα εκατόν σαράντα (140) είναι Subject (S), τα εκατόν έξι (106) είναι Object (O) και τα επτακόσια πενήντα εννέα (759) δεν είναι τίποτα από τα δύο (None).



Εικόνα 3: Εφαρμογή πειραμάτων στο πρόγραμμα RapidMiner.

4.3 Αποτελέσματα πειραμάτων

Ένα πολύ χρήσιμο και διαισθητικό μέτρο είναι το confusion matrix, το οποίο είναι χτισμένο από τη λίστα των προβλεπόμενων τάξεων έναντι των πραγματικών τάξεων. Πόσο είναι το ποσοστό κατά το οποίο η προβλεπόμενη τάξη συνάδει με την πραγματική τάξη;

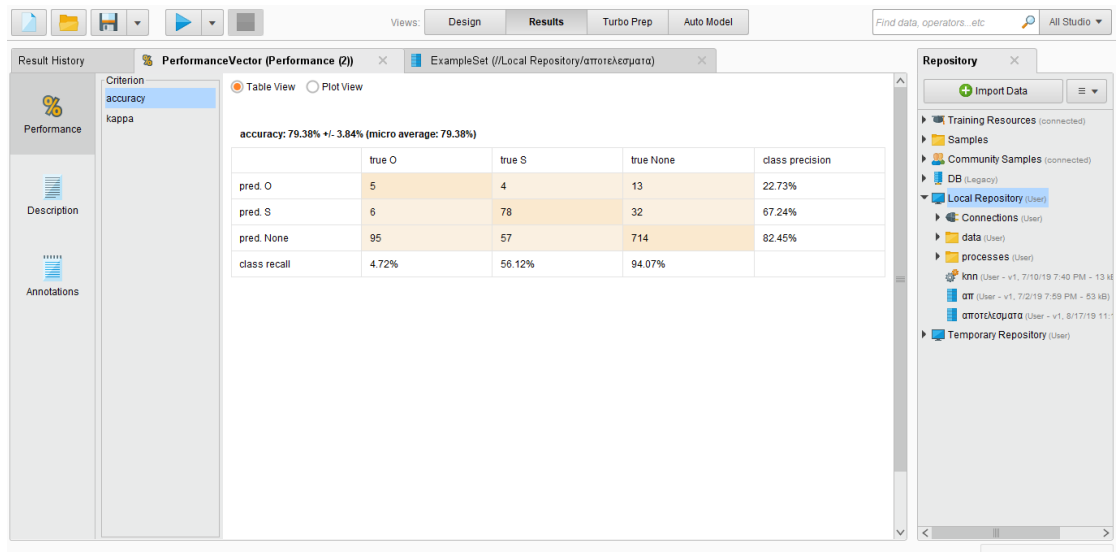
Precision : Ακρίβεια είναι ο αριθμός της σωστής πρόβλεψης διαιρούμενος με τον αριθμό των συνολικών προβλέψεων που έγιναν. Διαισθητικά, μια υψηλή ακρίβεια για μια τάξη σημαίνει ότι αν τα μοντέλα μας προβλέπουν αυτή την τάξη, είναι πολύ πιθανό να είναι αλήθεια. Ένα μοντέλο υψηλής ακρίβειας θα είναι χρήσιμο σε εκείνες τις περιπτώσεις όπου πρέπει να έχουμε υψηλή εμπιστοσύνη στην πρόβλεψή μας (για παράδειγμα στην ιατρική διάγνωση).

Η ακρίβεια μπορεί να υπολογιστεί χωριστά για κάθε κατηγορία. Γραφικά, είναι το άθροισμα του στοιχείου στη διαγώνιο του confusion matrix διαιρούμενο με τον συνολικό αριθμό των προβλέψεων που έγιναν. Η ακρίβεια είναι ένα μέτρο που μπορεί να είναι χρήσιμο όταν το πρόβλημα έχει ισορροπημένες τάξεις (για παράδειγμα στην οπτική αναγνώριση χαρακτήρων) και θέλουμε να δώσουμε έμφαση στις ακριβείς αντιστοιχίες. Ευτυχώς η ακρίβεια υποφέρει από τα μη ισορροπημένα σύνολα δεδομένων, ένα τυπικό παράδειγμα είναι η ανάκτηση πληροφοριών.

Recall : Ανάκληση είναι ο αριθμός των σωστών προβλέψεων διαιρούμενος με τον συνολικό αριθμό των στοιχείων που υπάρχουν σε αυτή την κατηγορία. Γραφικά, είναι η τιμή στη διαγώνιο, διαιρούμενη με το άθροισμα των τιμών στη σειρά. Αν η ανάκληση είναι υψηλή, αυτό σημαίνει ότι τα μοντέλα μας καταφέρνουν να ανακτήσουν τα περισσότερα στιγμιότυπα αυτής της κατηγορίας. Η απόκτηση της υψηλής ανάκλησης είναι πολύ εύκολη, αρκεί να πούμε ότι όλα αντιστοιχούν στην τάξη και μπορείτε να είστε σίγουροι ότι όλα τα στοιχεία θα ανακτηθούν.

Στο confusion matrix παρουσιάζεται ο αριθμός των προβλέψεων σε σχέση με τις πραγματικές τιμές τους. Στις γραμμές σημειώνονται οι τρεις τιμές των προβλέψεων (S, O και None), ενώ στις στήλες οι τρεις πραγματικές τιμές των λέξεων (S, O και None αντίστοιχα). Συμπληρωματικά, στο σύμπλεγμα των γραμμών και των στηλών παρουσιάζεται ο αριθμός των λέξεων που έχουν την αντίστοιχη τιμή πρόβλεψης στην γραμμή και την αντίστοιχη πραγματική τιμή τους στην στήλη. Για παράδειγμα, στην πρώτη γραμμή και την πρώτη στήλη φαίνεται ο αριθμός των λέξεων που έχουν προβλέψιμη και πραγματική τιμή O, στην πρώτη γραμμή και την δεύτερη στήλη φαίνεται ο αριθμός των λέξεων που έχουν προβλέψιμη τιμή O και πραγματική τιμή S, στην πρώτη γραμμή και την τρίτη στήλη φαίνονται πόσες λέξεις είχαν πρόβλεψη O και τελικά έχουν τιμή None και τα λοιπά.

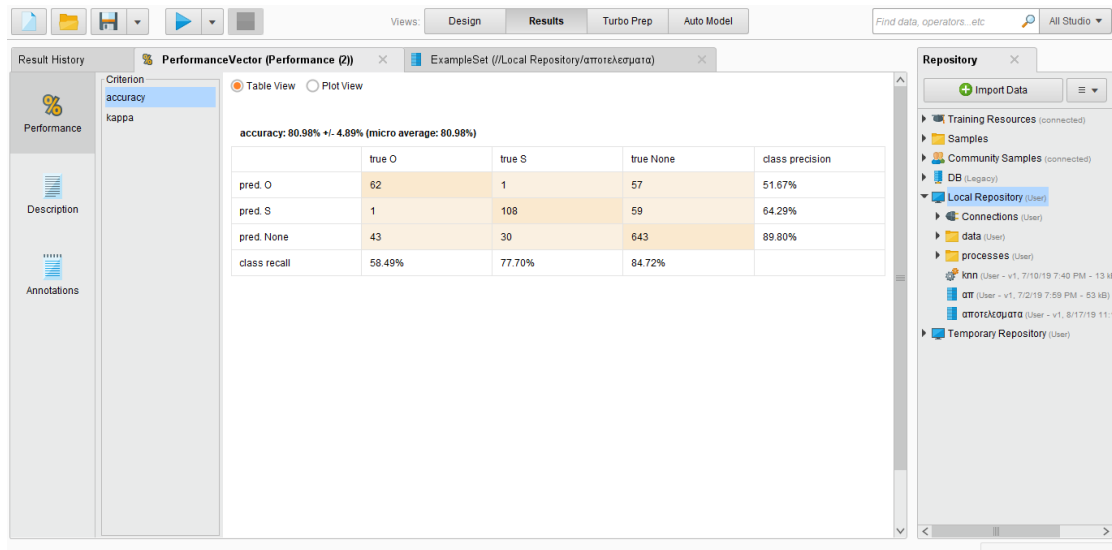
ID3



Εικόνα 4: Αποτέλεσμα πειράματος με τον id3.

Στο confusion matrix φαίνονται τα αποτελέσματα της έρευνας για τον id3. Οι προβλέψιμες τιμές O είναι και οι πραγματικές σε 5 λέξεις. Οι προβλέψιμες τιμές S είναι και οι πραγματικές σε 78 λέξεις. Οι προβλέψιμες τιμές None είναι και οι πραγματικές σε 714 λέξεις. Επομένως, οι προβλέψιμες τιμές συνάδουν με τις πραγματικές σε 797 από τα 1005 παραδείγματα μηχανικής μάθησης. Η ακρίβεια αυτού του αλγορίθμου είναι στο ποσοστό 79,38%. Επίσης, φαίνονται στον πίνακα και τα ποσοστά ακρίβειας και ανάκλησης της κάθε τάξης.

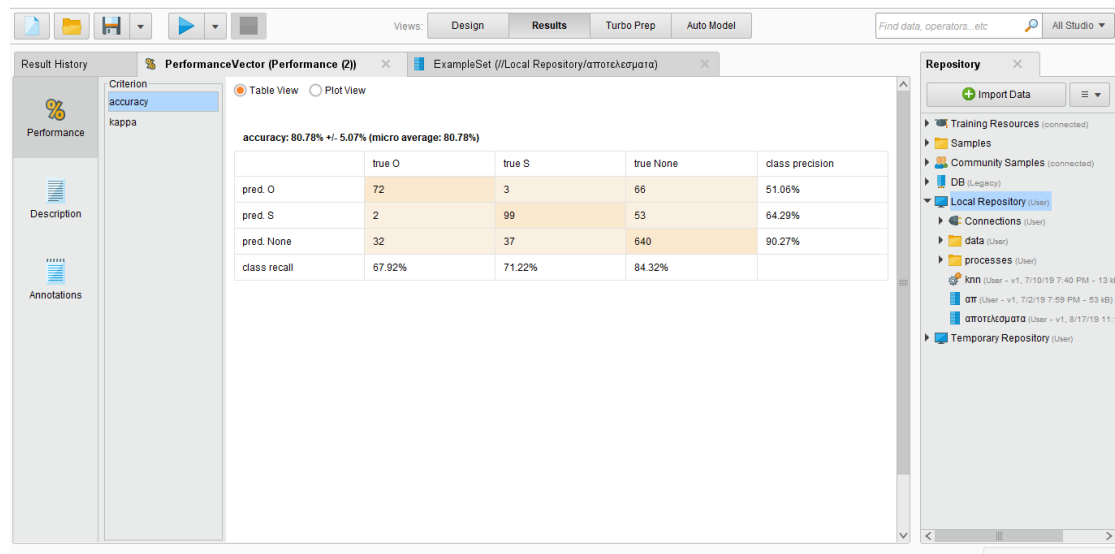
K-NN



Εικόνα 5: Αποτέλεσμα πειράματος με τον πλησιέστερο γείτονα.

Στο confusion matrix φαίνονται τα αποτελέσματα της έρευνας για τον k-nn. Οι προβλέψιμες τιμές O είναι και οι πραγματικές σε 62 λέξεις. Οι προβλέψιμες τιμές S είναι και οι πραγματικές σε 108 λέξεις. Οι προβλέψιμες τιμές None είναι και οι πραγματικές σε 643 λέξεις. Επομένως, οι προβλέψιμες τιμές συνάδουν με τις πραγματικές σε 813 από τα 1005 παραδείγματα μηχανικής μάθησης. Η ακρίβεια αυτού του αλγορίθμου είναι στο ποσοστό 80,98%. Επίσης, φαίνονται στον πίνακα και τα ποσοστά ακρίβειας και ανάκλησης της κάθε τάξης.

NAÏVE BAYES



Εικόνα 6: Αποτέλεσμα πειράματος με τον naïve bayes.

Στο confusion matrix φαίνονται τα αποτελέσματα της έρευνας για τον naïve bayes. Οι προβλέψιμες τιμές O είναι και οι πραγματικές σε 72 λέξεις. Οι προβλέψιμες τιμές S είναι και οι πραγματικές σε 99 λέξεις. Οι προβλέψιμες τιμές None είναι και οι πραγματικές σε 640 λέξεις. Επομένως, οι προβλέψιμες τιμές συνάδουν με τις πραγματικές σε 811 από τα 1005 παραδείγματα μηχανικής μάθησης. Η ακρίβεια αυτού του αλγορίθμου είναι στο ποσοστό 80,78%. Επίσης, φαίνονται στον πίνακα και τα ποσοστά ακρίβειας και ανάκλησης της κάθε τάξης.

Ως γενικό συμπέρασμα, είναι εύκολο να αναγνωριστούν τα υποκείμενα, καθώς φανερώνονται από το ρήμα της πρότασης. Το δύσκολο είναι αναγνωριστούν τα αντικείμενα, επειδή πολλές φορές είναι ολόκληρες ονοματικές φράσεις ή ακόμη και δευτερεύουσες προτάσεις. Γενικότερα, οι αλγόριθμοι έχουν αρνητικά αποτελέσματα σε tweets, τα οποία δεν έχουν υποκείμενο. Σε αυτά τα tweets το υποκείμενο υπονοείται εύκολα από το ρήμα, αλλά ο αλγόριθμος δεν μπορεί να το αναγνωρίσει καθώς ως μηχανή μάθησης δεν μπορεί να αναγνωρίσει λέξη ενώ αυτή δεν υπάρχει.

Το ρήμα της πρότασης, οι λέξεις που προηγούνται και αυτές που έπονται τη λέξη που ερευνάται, καθώς και τα μέρη του λόγου τους, είναι τα πιο χρήσιμα στοιχεία για την αναγνώριση του υποκειμένου και του αντικειμένου. Ασήμαντα χαρακτηριστικά μπορούν να θεωρηθούν ο αριθμός των κομμάτων και ο αριθμός άλλων ρημάτων που μεσολαβούν μεταξύ του κύριου ρήματος και της λέξης που ερευνάται. Αυτά τα χαρακτηριστικά δεν είναι απαραίτητα για την αναγνώριση του υποκειμένου και του αντικειμένου, παρά μόνο για τον διαχωρισμό κύριων και δευτερευουσών προτάσεων, καθώς και για την αναγνώριση κύριων και βοηθητικών ρημάτων. Επιπλέον, παραπλανητική μπορεί να θεωρηθεί η απόσταση της λέξης που ερευνάται από το ρήμα, διότι το υποκείμενο μπορεί να βρίσκεται μετά το ρήμα και, όχι απαραίτητα πριν από αυτό, ενώ το αντικείμενο μπορεί να βρίσκεται σε οποιαδήποτε θέση της πρότασης. Τέλος, ένα χαρακτηριστικό το οποίο δεν βοηθάει καθόλου στην έρευνα είναι το μέρος του λόγου του ρήματος, καθώς ο αριθμός και ο χρόνος του ρήματος δεν παίζει κανένα ρόλο στην αναγνώριση του υποκειμένου και του αντικειμένου.

4.4 Συμπεράσματα πειραμάτων

Εν κατακλείδι, η γνώση είναι αυτή που αναδεικνύει την σημαντικότητα της πληροφορίας συσχετίζοντας τη με χρήσιμα συμπεράσματα ή αναγκαίες ενέργειες. Η αναπαράσταση, ωστόσο, της γνώσης χρειάζεται για να εξάγονται αυτά τα συμπεράσματα από τον υπολογιστή με αυτόματο τρόπο.

Ως συμπέρασμα των πειραμάτων της έρευνας που πραγματοποιήθηκε, ο πλησιέστερος γείτονας (k-nn) έχει καλύτερη απόδοση από τους άλλους δύο. Τα ποσοστά ακρίβειας και ανάκλησης των αποτελεσμάτων του είναι μεγαλύτερα από τους Naïve Bayes και id3 αλγόριθμους. Αυτό σημαίνει ότι οι προβλεπόμενες τιμές των λέξεων, αυτές δηλαδή που προβλέπονται από τις παρατηρήσεις και το μοντέλο, είναι και οι πραγματικές τιμές τους. Ωστόσο, ο αλγόριθμος Naïve Bayes έχει εξίσου καλά αποτελέσματα με τον αλγόριθμο των πλησιέστερων γειτόνων k-nn. Οι δύο αυτοί αλγόριθμοι έχουν υψηλά ποσοστά ακρίβειας και ανάκλησης. Επιπλέον, έχουν μικρή έως και απειροελάχιστη απόκλιση στα αποτελέσματά τους. Ο αλγόριθμος id3 δεν αποδίδει αποτελεσματικά στην ταξινόμηση των λέξεων όσο οι άλλοι δύο. Συνεπώς, η πρόβλεψη και η ταξινόμηση των λέξεων συνάδουν κατά μεγάλο ποσοστό στην εκτέλεση των k-nn και naïve bayes αλγορίθμων.

Ο αλγόριθμος Naive Bayes είναι ένας αλγόριθμος στον οποίο κάθε χαρακτηριστικό που ταξινομείται είναι ανεξάρτητο το ένα από το άλλο. Η ύπαρξη λοιπόν πολλών χαρακτηριστικών στην έρευνα φέρνει σύγχυση στον αλγόριθμο και συνεπώς, αποτελεί πρόβλημα στην ταξινόμηση των λέξεων και τα αποτελέσματά του δεν έχουν μεγάλο ποσοστό ακρίβειας. Ο αλγόριθμος των k-πλησιέστερων γειτόνων δεν φέρνει εκατό τοις εκατό ακριβή αποτελέσματα, ίσως επειδή η λειτουργία του περιορίζεται τοπικά, στους πλησιέστερους γείτονες και συνεπώς, δεν μπορεί να έχει ως δεδομένα ένα ευρύ φάσμα χαρακτηριστικών. Ωστόσο, σημαντικό ρόλο παίζει και το ασύντακτο σώμα κειμένων. Η ακρίβεια του ID3 μπορεί να αυξηθεί ως ποσοστό με την καλύτερη προεπεξεργασία των δεδομένων, το οποίο είναι δύσκολο για ένα σώμα κειμένων συλλεγμένο από το Twitter.

Συμπερασματικά, η εκτέλεση των πειραμάτων στους αλγόριθμους αντιμετωπίζει προβλήματα και συνεπώς, τα αποτελέσματά τους δεν είναι ακριβή. Αυτό οφείλεται κατά κύριο λόγο στο σώμα κειμένων, το οποίο έχει συνταχθεί από συλλογή των tweets. Αυτό το σώμα κειμένων δεν είναι ορθά συνταγμένο, καθώς χρησιμοποιείται πολλές φορές στα tweets η αργκό, μια συνηθισμένη γλώσσα που είναι συνηθισμένη από τους ανθρώπους σε κάθε περιοχή του κόσμου. Επιπλέον, στα tweets παρατηρείται πολύ συχνά και το φαινόμενο της ερωτοαπάντησης, κατά το οποίο δεν διακρίνεται εύκολα το υποκείμενο ή το αντικείμενο και οι θέσεις αυτών. Σε αντίθεση με αυτό το σώμα κειμένων, τα επιστημονικά κείμενα έχουν ορθά συντακτικά φαινόμενα και για αυτό το λόγο, είναι κατάλληλα για την ανάλυση και ταξινόμηση λέξεων και μερών του λόγου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Androutsopoulos, J., & Beißwenger, M. (2008). Introduction: Data and methods in computer-mediated discourse analysis. *Language@ internet*, 5(2).

Bergs, A. (2006). Analyzing online communication from a social network point of view: Questions, problems, perspectives. *Language@ internet*, 3(3).

Bharti, S. K., Babu, K. S., & Jena, S. K. (2015, August). Parsing-based sarcasm sentiment recognition in twitter data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 1373-1380). ACM.

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR mental health*, 3(2).

Carnie, A. (2013). *Syntax: A generative introduction*. John Wiley & Sons.

Ch, R., & MAP, M. *Bayesian Learning*.

Daelemans, W., Buchholz, S., & Veenstra, J. (1999). Memory-based shallow parsing. *arXiv preprint cs/9906005*.

Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011.

Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., ... & Van Genabith, J. (2011). # hardtoparse: POS Tagging and Parsing the Twittiverse. In *AAAI 2011 workshop on analyzing microtext* (pp. 20-25).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... & Smith, N. A. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments. *Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science*.

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Entropy*, 17, 252.

Goldin, C. D., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, MA: Belknap Press of Harvard University Press.

Goutsos, D., Fragkaki, G., Γούτσος, Δ., & Φραγκάκη, Γ. (2015). Σώματα κειμένων και γλωσσολογία σωμάτων κειμένων.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive science*, 41(4), 976-995.

Holton, D., P. Mackridge & I. Φιλippάκη-Warburton (1999). *Γραμματική της ελληνικής γλώσσας*. (Μτφρ. Β. Σπυρόπουλος). Αθήνα: Πατάκης.

Ide, N. (1998, May). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In Proceedings of the First International Language Resources and Evaluation Conference (pp. 463-70).

Ide, N., & Véronis, J. (1994, August). MULTEXT: Multilingual text tools and corpora. In Proceedings of the 15th conference on Computational linguistics-Volume 1 (pp. 588-592). Association for Computational Linguistics.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 151-160). Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2014). Speech and language processing (Vol. 3). London: Pearson.

Kennedy, G. (2014). An introduction to corpus linguistics. Routledge.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). A dependency parser for tweets. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1001-1012).

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archana Bhatia, Chris Dyer, and Noah A. Smith. 2014. A Dependency Parser for Tweets. In Proc. of EMNLP.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., & Smith, N. A. (2018). Parsing Tweets into Universal Dependencies. arXiv preprint arXiv:1804.08228.

Lyons, J. (1995). Εισαγωγή στη γλωσσολογία. Μ. Αραποπούλου κ. ά., Καρανάσιος Γ.(επιμ.). Πατάκης: Αθήνα.

Magoulas, G., Magoula, E., Μαγουλάς, Γ., & Μαγουλά, Ε. (2017). Νεότερη Γλωσσολογία.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330.

Muggleton, S., Bain, M., Hayes-Michie, J., & Michie, D. (1989). An experimental comparison of human and machine learning formalisms. In Proceedings of the sixth international workshop on Machine learning (pp. 113-118).

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

Pantazara, A. V., Mantzari, E., Πανταζάρα, Α. Β., & Μάντζαρη, Ε. (2015). Αναζήτηση σε σώματα κειμένων.

Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proc. of ACL.

- Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.
- Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of memory and language*, 57(3), 348-379.
- Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. " O'Reilly Media, Inc."
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
- Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(1), 178-183.
- Sang, E. F. (2002). Memory-based shallow parsing. *Journal of machine learning research*, 2(Mar), 559-594.
- Sato, S., Matsuyoshi, S., & Kondoh, Y. (2008, May). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *LREC*.
- Stuart, K., & Botella, A. (2009). Lingüística de corpus, análisis de redes y matrices de co-ocurrencias. Departamento de Lingüística Aplicada, Universitat Politècnica de València.(612-30p).
- Swalin, A. (2018). *Choosing the Right Metric for Evaluating Machine Learning Models—Part 2*.
- Swayamdipta, L. K. N. S. S. A Dependency Parser for Tweets.
- Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing Tweets into Universal Dependencies. In *Proc. of NAACL*.
- Zanzotto, F. M., Pennacchiotti, M., & Tsioutsoulouklis, K. (2011, July). Linguistic redundancy in twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 659-669). Association for Computational Linguistics.
- Βλαχάβας Ι., & Κεφαλάς, Π., & Βασιλειάδης, Ν., & Κόκκορας, Φ., & Σακελλαρίου, Η. Προηγμένη Διασύνδεση με το Περιβάλλον, Τεχνητή Νοημοσύνη.
- Δήμας, Α. (2013). *Τεχνικές για την εξαγωγή γνώσης από την πλατφόρμα του Twitter* (Doctoral dissertation).
- Ιωαννίδου, Ε. Η διδασκαλία της γραμματικής: Από τη γλωσσική δομή στο κειμενικό είδος. Τμήμα Επιστημών της Αγωγής, Πανεπιστήμιο Κύπρου.
- Κερμανίδου, Κ. Α. (2005). *Αυτόματη μάθηση συντακτικών εξαρτήσεων και ανάπτυξη γραμματικών της ελληνικής γλώσσας*.

Κλαίρης, Χ., Μόζερ, Α., & Μπαμπινιώτης, Γ. (1996). Γραμματική της νέας ελληνικής, δομολειτουργική, επικοινωνιακή. Εκδ." Ελληνικά γράμματα".

Μηταρά, Ε. (2017). Τεχνικές εξαγωγής και ανάλυσης συναισθημάτων σε μηνύματα του Twitter και υλοποίηση της ανάλυσης σε περιβάλλον cloud (Doctoral dissertation).

Μπαζιώτης, Χ. (2017). Ανάλυση συναισθήματος στο twitter με βαθιά νευρωνικά δίκτυα (Master's thesis, Πανεπιστήμιο Πειραιώς).

Παϊδούση, Ε. (2016). Δέντρα αποφάσεων (Doctoral dissertation).

Παπασωτηρόπουλος, Γ. (2017). Υπολογιστικά ζητήματα για γράφους του Twitter. Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

Ποιμενίδου, Μ. (2008). Ο παιδαγωγικός λόγος στο νηπιαγωγείο. Μελέτη των αποκλίσεων στην οργάνωση των κειμενικών ειδών και τη συγκρότηση του νοήματος.

Ρούσσου, Ά. (2015). Από τη λέξη στη φράση. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Ρούσσου, Ά. (2015). Γλώσσα και γραμματική. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Ρούσσου, Ά. (2015). Γραμματικές λειτουργίες και (κενά) ορίσματα. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Ρούσσου, Ά. (2015). Συμφωνία και σχέσεις ορισμάτων. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Ρούσσου, Ά. (2015). Συντακτικές κατηγορίες. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Ρούσσου, Ά. (2015). Σύνταξη και λεξικό. Τμήμα φιλολογίας, Πανεπιστήμιο Πατρών.

Χαλάτσης, Κ. Αλγόριθμοι-Προγραμματισμός, Εισαγωγή στην Επιστήμη της Πληροφορικής και των Τηλεπικοινωνιών. Πανεπιστήμιο Αθηνών.

Χατζηδάκη, Α. (2009-2010). Η γλώσσα ως σύστημα και ως χρήση.

Χριστοπούλου, Α. (2017). Ανάλυση συναισθήματος με χρήση τεχνικών μηχανικής μάθησης (Doctoral dissertation).