ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
με θέμα:

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ ΑΥΤΟΜΑΤΗΣ ΑΝΑΓΝΩΡΙΣΗΣ ΓΕΩΓΡΑΦΙΚΟΥ ΙΔΙΩΜΑΤΙΣΜΟΥ ΤΟΥ ΣΥΓΓΡΑΦΕΑ ΣΕ ΣΥΛΛΟΓΗ ΚΕΙΜΕΝΩΝ ΑΠΟ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Σιμάκης Παναγιώτης
ΑΜ. 5227

Επιβλέπων Καθηγητής:
Μεγαλοοικονόμου Βασίλειος

Πάτρα, Σεπτέμβριος 2016

Πανεπιστήμιο Πατρών, Τμήμα Μηχανικών Η/Υ και Πληροφορικής
Παναγιώτης Σιμάκης

Ευχαριστίες

Στα πλαίσια της διπλωματικής, θα ήθελα να ευχαριστήσω μερικούς ανθρώπους για την καθοριστική συμβολή τους στην ολοκλήρωσή της.

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή κ. Μεγαλοοικονόμου Βασίλειο ο οποίος υπήρξε επιβλέποντας αυτής της εργασίας. Οι γνώσεις καθώς και οι συμβουλές του υπήρξαν καθοριστικές για την ολοκλήρωση της εργασίας.

Ακόμα ένα μεγάλο ευχαριστώ στην Δρ. Βασιλική Σιμάκη για την καθοδήγηση, τις συμβουλές και τη βοήθεια της καθ' όλο το διάστημα της υλοποίησης και της συγγραφής αυτής της εργασίας. Χωρίς την δικιά της καθοδήγηση το αποτέλεσμα της εργασίας δεν θα ήταν το ίδιο.

Τέλος η καθοριστική ήταν η στήριξη της οικογένειάς μου, των γονιών μου, της αδερφής μου και της Σοφίας, χωρίς αυτούς δεν θα είχα καταφέρει τίποτα.

Περίληψη

Η ραγδαία εξάπλωση των μέσων κοινωνικής δικτύωσης δημιουργεί όλο και περισσότερα ζητήματα προς διερεύνηση και μελέτη στην επιστημονική κοινότητα. Ο τεράστιος όγκος πληροφορίας από μόνος του αποτελεί πρόκληση ως προς τη διαχείριση του. Η οργάνωση της πληροφορίας βάσει θέματος, συγγραφέα, ηλικίας, φύλου και γεωγραφικής προέλευσης αποτελούν παραδείγματα προβλημάτων που επιζητούν λύση.

Αντικείμενο της συγκεκριμένης διπλωματικής είναι η ανάπτυξη μεθοδολογίας αυτόματης αναγνώρισης γεωγραφικού ιδιωματισμού του συγγραφέα μέσα από συλλογή κειμένων από μέσα κοινωνικής δικτύωσης. Αρχικά γίνεται βιβλιογραφική αναζήτηση στα πεδία της κατηγοριοποίησης κειμένου, της εξόρυξης γνώσης από κείμενο και της αυτόματης αναγνώρισης συγγραφέα. Στην συνέχεια προχωράμε στην συλλογή των δεδομένων τα οποία προέρχονται από μέσα κοινωνικής δικτύωσης και από χρήστες για τους οποίους μπορούνε να αντλήσουμε δημογραφικές πληροφορίες. Αφού γίνει η συλλογή δεδομένων με κείμενα από τα μέσα κοινωνικής δικτύωσης γίνεται προεπεξεργασία και επισημείωση των κειμένων με σκοπό στην συνέχεια να γίνει εξαγωγή χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών γίνεται βάσει γλωσσικών στοιχείων αλλά και βάσει ιδιωματισμών που προδίδουν τη γεωγραφική καταγωγή του συγγραφέα. Τέλος πραγματοποιούμε πειράματα κατηγοριοποίησης με τη χρήση αρκετών αλγορίθμων κατηγοριοποίησης συγκρίνοντας και αξιολογώντας ανάλογα τα αποτελέσματα που λαμβάνουμε.

Περιεχόμενα

Λίστα Πινάκων

Λίστα Εικόνων

Κεφάλαιο 1 – Εισαγωγικά

1 Γενική Περιγραφή

Η εξάπλωση του διαδικτύου τα τελευταία χρόνια έχει πάρει τεράστιες διαστάσεις. Η σχέση των ανθρώπων με το διαδίκτυο και ακόμη περισσότερο με τα κοινωνικά δίκτυα γίνεται καθημερινή και έντονη. Πλέον ο άνθρωπος διαλέγει τα κοινωνικά δίκτυα για να εκφραστεί σε μεγάλο βαθμό καθώς μέσα από αυτά σχολιάζει συμπεριφορές, εμφανίσεις, μουσικές, και καθετί που βρίσκει ενδιαφέρον. Αυτό έχει σαν αποτέλεσμα να χρησιμοποιείται ο γραπτός λόγος και ο όγκος που παράγεται να είναι πραγματικά τεράστιο. Έτσι λοιπόν προκύπτει η ανάγκη της ανάλυσης των γλωσσικών επιλογών των χρηστών. Αυτό οδηγεί και στην αναγνώριση του προφίλ του συγγραφέα.

Τα χαρακτηριστικά με τα οποία μπορούμε να χωρίζουμε τους συγγραφείς είναι αρκετά, όπως το φύλο, η καταγωγή, το μορφωτικό επίπεδο, το επάγγελμα καθώς και η ηλικία. Στην παρούσα διπλωματική εστιάζουμε στον διαχωρισμό του συγγραφέα βάσει του γεωγραφικού ιδιωματισμού του. Στο συγκεκριμένο χαρακτηριστικό δεν υπάρχει αρκετό ερευνητικό υλικό καθώς μέχρι στιγμής οι έρευνες έχουν περιοριστεί στο φύλο και την ηλικία του συγγραφέα.

2 Στόχοι της Εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η διαμόρφωση μια συνολικής μεθοδολογίας εξειδικευμένης στην αυτόματη αναγνώριση συγγραφέα βάσει του γεωγραφικού ιδιωματισμού. Δηλαδή έχοντας ως είσοδο ένα σύνολο κειμενο, το σύστημα αυτό να είναι σε θέση να δώσει στην έξοδο την γεωγραφική προέλευση του συγγραφέα.

Για την κατηγοριοποίηση απαιτείται ένα σύνολο χαρακτηριστικών (features), μέσω των οποίων κάθε κείμενο θα μετατραπεί σε ένα διάνυσμα χαρακτηριστικών. Στην βιβλιογραφία έχει προταθεί ένα μεγάλο πλήθος χαρακτηριστικών. Στόχος μας είναι τα χαρακτηριστικά που επιλέγονται να είναι οσο το δυνατόν πιο ανεξάρτητα από τη γλώσσα του κειμένου έτσι ώστε να μπορεί να εφαρμοστεί σε ένα μεγάλο αριθμό διαφορετικών κειμένων.

3 Δομή Εργασίας

Σε αυτό το σημείο θα περιγραφεί η δομή της εργασίας.

Στο Κεφάλαιο 2 παρουσιάζονται τα επιστημονικά πεδία της επεξεργασίας φυσικής γλώσσας όπως η κατηγοριοποίησης κειμένου καθώς και η εξαγωγή προφίλ συγγραφέα (Authorship attribution) αλλά και συνολικά τα επιστημονικά πεδία που συνδέονται ευρύτερα όπως η μηχανική μάθηση, η εξόρυξη γνώσης.

Στο Κεφάλαιο 3 περιγράφεται η μεθοδολογία που ακολουθήθηκε, και στο δεύτερο μέρος του κεφαλαίου γίνεται μια συνοπτική περιγραφή των εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση των πειραμάτων.

Στο Κεφάλαιο 4 γίνεται μια αναλυτική περιγραφή της πειραματικής διαδικασίας. Αναλύεται διεξοδικά η διαδικασία συλλογής των κειμένων, η προεπεξεργασία, η εξαγωγή των χαρακτηριστικών, τα πειράματα κατηγοριοποίησης, και τα αποτελέσματα.

Στο Κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα που προκύπτουν μετά την ολοκλήρωση της εργασίας, τα ανοιχτά θέματα προς μελλοντική διερεύνηση καθώς και τις δυσκολίες που προέκυψαν στην παρούσα εργασία.

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο

1 Επεξεργασία Φυσικής Γλώσσας

Η ραγδαία ανάπτυξη των τηλεπικοινωνιών, των πληροφοριακών συστημάτων και των υπολογιστών έχει επιφέρει συνέπειες στον τρόπο που δουλεύουμε, που επικοινωνούμε, διασκεδάζουμε και αγοράζουμε προϊόντα. Φυσικό επακόλουθο η δημιουργία τεράστιας ποσότητας πληροφορίας, μεγάλο μέρος της οποίας είναι σε μορφή κειμένου. Η συγκεκριμένη μορφή πληροφορίας διαφέρει από τις παραδοσιακές μορφές δεδομένων με τις οποίες έχουμε μάθει να δουλεύουμε όλα αυτά τα χρόνια. Αυτό οφείλεται στην φυσικότητα του γραπτού κειμένου, καθώς είναι κάτι αφηρημένο και αδόμητο, για αυτό το λόγο δημιουργούνται δυσκολίες τις οποίες καλείται να δώσει λύσεις το πεδίο της επεξεργασίας φυσικού κειμένου.

Γενικά η διαδικασία της επεξεργασίας φυσικού κειμένου προσπαθεί να αποτελέσει τον κρίκο ανάμεσα στον άνθρωπο όπως απλό χρήστη και τα υπολογιστικά συστήματα (Αθανασοπούλου, 2006). Στόχος είναι η διευκόλυνση της επικοινωνίας των χρηστών με τα υπολογιστικά συστήματα. Ο στόχος αυτός επιτυγχάνεται μέσα από την πληροφορία που εξάγεται από το φυσικό κείμενο.

Τα τελευταία χρόνια για την αξιοποίηση των διαδικασιών επεξεργασίας φυσικής γλώσσας έχουν προταθεί μια σειρά εφαρμογών που προσεγγίζουν μεγάλο εύρος χρηστών. Μερικές από αυτές είναι:

- Μηχανική Μετάφραση (Machine Translation)
- Αναγνώριση προφορικού λόγου (Speech Recognition)
- Ανάκτηση πληροφορίας (Information Retrieval)
- Περίληψη κειμένων (Summarization)

Οι παραπάνω μέθοδοι χρησιμοποιούνται ευρέως για να εξαχθεί πληροφορία από τεράστιες συλλογές κειμένων που είναι διαθέσιμα στο διαδίκτυο. Για την εξόρυξη γλωσσολογικής πληροφορίας ο τεράστιος όγκος κειμένων αποτελεί σημαντικό στοιχείο. Η γλωσσολογική πληροφορία στην συνέχεια χρησιμοποιείται στην βελτίωση των συστημάτων επεξεργασίας φυσικής γλώσσας.

Σε κάθε περίπτωση το πλήθος των αδόμητων κειμένων που είναι διαθέσιμα στο διαδίκτυο και περιέχουν πληροφορία είναι αυτά που δίνουν κίνητρο για περαιτέρω έρευνα στο πεδίο της μεθοδολογίας της επεξεργασίας φυσικού κειμένου.

Οι περισσότερες εφαρμογές επεξεργασίας φυσικής γλώσσας απαιτούν κωδικοποιημένη γνώση η οποία μπορεί να αποκτηθεί με χρήση τυπικών μεθόδων επεξεργασίας κειμένων. Η γνώση αυτή, είναι χρήσιμη πληροφορία που έχει να κάνει με τα μέρη του λόγου, τη σημασία των λέξεων, τη φωνητική, τη γραμματική καθώς και τη δομή του κειμένου κλπ καθώς και όποια άλλη πληροφορία περιέχεται σε ένα κείμενο. Στην περίπτωση μικρών εφαρμογών, η γνώση αυτή μπορεί να εισαχθεί

χειροκίνητα, για εφαρμογές όμως γενικού σκοπού το πλήθος της απαιτούμενης γνώσης καθιστά την παραπάνω μέθοδο μη εφαρμόσιμη, και προσανατολιζόμαστε σε αυτόματες μεθόδους.

Έτσι δημιουργείται η ανάγκη για εφαρμογές επεξεργασίας φυσικής γλώσσας με αυτόματη ή ήμι-αυτόματη απόκτηση γνώσης μια σειρά τεχνικών που καλύπτονται με τον όρο «Στατιστική Επεξεργασία Φυσικής Γλώσσας». Παρότι με τον συγκεκριμένο όρο φαίνεται να αποκλείονται προσεγγίσεις που δεν χρησιμοποιούν υπολογισμούς στατιστικής συχνότητα ή βασικές αρχές θεωρίας πιθανοτήτων, χρησιμοποιείται σε μεγάλο βαθμό καθώς είναι περιεκτικός. Όμως αντί του παραπάνω όρου θα μπορούσαμε να χρησιμοποιήσουμε τον όρο «Αυτόματη ή Ήμι-αυτόματη Απόκτηση Γνώσης από Γλωσσολογικές Πηγές».

Η διαδικασία της επεξεργασίας φυσικής γλώσσα αποτελείται από διάφορες μεθόδους σημασιολογικής και συντακτικής ανάλυσης εγγράφων. Για την υλοποίηση των κανόνων αυτών απαιτείται η δημιουργία συνόλου κανόνων είτε με το χέρι είτε μέσω αυτόματων διαδικασιών αυτόματης εκπαίδευσης των μεθόδων που χρησιμοποιούνται για την επεξεργασία φυσικού κειμένου. Το πλεονέκτημα των συνόλων κανόνων που δημιουργούνται είναι η εύκολη επέκταση τους αλλά και τις μικρές απαιτήσεις γλωσσολογικών γνώσεων του προς επεξεργασία κειμένου. Όμως η απόδοση των εκπαιδεύσιμων συνόλων συχνά δεν είναι τόσο καλή σε σχέση με την δημιουργία κανόνων χειρωνακτικά.

2 Βασικές Έννοιες Εξόρυξης Γνώσης

Η εξόρυξη γνώσης ή Εξόρυξη Δεδομένων με βάσει την αυστηρή μετάφραση του Data Mining αποτελεί την διαδικασίας εξεύρεσης πληροφορίας από μεγάλες βάσεις δεδομένων με τη χρήση κατάλληλων τεχνικών/αλγορίθμων (Data Mining, Wikipedia). Αν θέλαμε να αναπαραστήσουμε τις «ρίζες» της εξόρυξης γνώσης/δεδομένων ως πεδίο θα έμοιαζε έτσι:

2.1 Συσταδοποίηση (Clustering)

Η διαδικασία κατά την οποία ένα σύνολο από τα «αντικείμενα» διαχωρίζονται σε ένα σύνολο από λογικές μονάδες. Όταν έχουμε καταχώρηση αντικείμενων στην ίδια ομάδα τότε αυτό αποτυπώνεται ως ομοιότητα των αντικειμένων και αντίστροφα. Η ομοιότητα ή μη των αντικειμένων εξαρτάται ουσιαστικά από την μορφή των «αντικειμένων» αλλά και το πρόβλημα καθαυτό. Ο ορισμός της αυστηρής συσταδοποίησης είναι:

Δοθέντος ενός συνόλου διανυσμάτων $X=\{x_1, x_2, \ldots x_n\}$ ζητούνται m σύνολα-ομάδες $C_1, C_2 \ldots C_m$, με m<<n έτσι ώστε :

$C_i \neq \emptyset \; \forall \, i=1,2,3,\ldots m$ και οι m ομάδες αποτελούν διαμεριση του συνόλου X.

2.2 Κανόνες Συσχέτισης (Association rule mining)

Είναι από τις νεότερες τεχνικές εξόρυξης γνώσης. Οι πληροφορίες μου συλλέγονται μπορούν να παράγουν σαν αποτέλεσμα ενδιαφέροντα πρότυπα και συσχετίσεις τα οποία μπορούν να βρουν εφαρμογή σε τομείς της καθημερινότητα του ανθρώπου. Ο ορισμός των κανόνων συσχέτισης δίνεται παρακάτω.

Έστω $I=\{i_1, i_2, \ldots, i_n\}$ ένα σύνολο από διακριτά στοιχεία, που αποκαλούνται items (αντικείμενα). Έστω ακόμα $D=\{t_1, t_2, \ldots t_m\}$ ένα σύνολο από δοσοληψίες (transactions), όπου κάθε δοσοληψία Τ είναι ένα σύνολο από αντικείμενα τα οποία ονομάζονται *itemset*, και όπου ισχύει $T \subset I$. Κάθε δοσοληψία ταυτίζεται με ένα μοναδικό αναγνωριστικό που καλείται TID.

Ένας κανόνας είναι μία συσχέτιση της μορφής $X \Rightarrow Y$ όπου $X \subset Y$, $Y \subset I$ και $X \cap Y = \emptyset$. Το πρώτο μέλος του κανόνα ονομάζεται *υπόθεση* ενώ το δεύτερο ονομάζεται *συμπέρασμα*.

Υπάρχουν δύο βασικές μετρικές στον συσχετισμό κανόνων. Η υποστήριξη(support) που σημαίνει ότι ο κανόνας $X \Rightarrow Y$ έχει υποστήριξη s, αν το s% των δοσοληψιών στο D περιέχουν το $(X \cup Y)$ και η εμπιστοσύνη που σημαίνει ότι ο κανόνας $X \Rightarrow Y$ ισχύει στο D, αν το c% των δοσοληψιών στο D που περιέχουν το X, περιέχουν επίσης και το Y. Άρα σύμφωνα και με το παράδειγμα του πίνακα η υποστήριξη για το στοιχειοσύνολο(itemset) {Milk, Bread,Butter} θα είναι s({Milk, Bread,Butter}) = 2/5 ενώ η υποστήριξη θα είναι c({Milk, Bread,Butter}) = 2/3.

Από τα παραπάνω προκύπτει ότι ο κανόνας $X \Rightarrow Y$ έχει υποστήριξη s, όταν $\sup(X \Rightarrow Y) = \sup(X \cup Y)$ και εμπιστοσύνη c, όταν $conf(X \Rightarrow Y) = supp(X \cup Y) / supp(X)$.

2.3 Κατηγοριοποίηση (Classification)

Αποτελεί και αυτή με την σειρά της τεχνική εξόρυξης γνώσης, σε αυτή τη διαδικασία κάθε στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο όρος στην βιβλιογραφία μπορεί να βρεθεί και ως ταξινόμηση. Στόχος της συγκεκριμένης διαδικασίας είναι η δημιουργία ενός μοντέλου το οποία με τη σειρά του στο μέλλον θα μπορεί να κατηγοριοποιεί νέα δεδομένα.

Η διαδικασία της κατηγοριοποίησης μπορεί να περιγραφεί από τα εξής δύο βήματα:

1. **Εκμάθηση (Learning)**: Σε αυτό το πρώτο βήμα με βάση ένα σύνολο προ-κατηγοριοποιημένων παραδειγμάτων δημιουργείται ή/και προσδιορίζεται το μοντέλο. Για το συγκεκριμένο βήμα τα παραδείγματα που χρησιμοποιούμε αποκαλούνται δεδομένα εκπαίδευσης (training set). Για να διαμορφωθεί το μοντέλο τα δεδομένα αναλύονται από ένα αλγόριθμο κατηγοριοποίησης. Επειδή τα δεδομένα εκπαίδευσης ανήκουν εκ των προτέρων σε μια κατηγορία, η οποία είναι γνωστή, η διαδικασία της κατηγοριοποίησης αποτελεί μια μέθοδο εποπτευόμενης μάθησης. Το μοντέλο (αλλιώς και κατηγοριοποιητής) αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων απόφασης ή μαθηματικών τύπων.

2. **Κατηγοριοποίηση (Classification)**: Το επόμενο βήμα μετά την δημιουργία του μοντέλου είναι η αξιολόγηση του. Για να το επιτύχουμε χρησιμοποιούμε τα δεδομένα ελέγχου (test set) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί και τα δεδομένα ελέγχου. Στην συνέχεια η κατηγορία που διαμορφώθηκε με βάση τα δεδομένα ελέγχου συγκρίνεται με την πρόβλεψη που έγινε με τα δοκιμαστικά δεδομένα, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται με βάση το ποσοστών των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπο εκπαίδευση μοντέλο.

Αν το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων

3 Κατηγοριοποίηση Κειμένου (Text Categorization)

3.1 Εισαγωγή

Η κατηγοριοποίηση κειμένου (ΚΚ – γνωστή και ως ταξινόμηση κειμένου) είναι η διαδικασία της αυτόματης ταξινόμησης εγγράφων σε κατηγορίες (ή κλάσεις ή θέματα) ενός προκαθορισμένου συνόλου (Sebastiani, 2002). Αυτή η διαδικασία βρίσκεται στο σταυροδρόμι της ανάκτησης πληροφορίας (ΑΠ) και της μηχανικής μάθησης (ΜΜ). Η ΚΚ παρουσιάζει ιδιαίτερο ενδιαφέρον τα τελευταία 10 χρόνια για ερευνητές και προγραμματιστές.

Οπουδήποτε υπάρχουν προκαθορισμένες κλάσεις, τα κείμενα κατηγοριοποιούνται με το χέρι από το χρήστη, ως εκ τούτου τα δεδομένα που παράγονται μπορούν να αξιοποιηθούν για την εκμάθηση της έννοιας οτι ο χρήστης αποδίδει γνωρίσματα στις κλάσεις, με αυτό τον τρόπο προσεγγίζουμε την ακρίβεια της ταξινόμησης, κάτι το οποίο θα ήταν αδιανόητο χωρίς αυτά τα δεδομένα.

Για τους ερευνητές της μηχανικής μάθησης, το ενδιαφέρον έγγειυται στο γεγονός οτι οι εφαρμογές ανάκτησης πληροφορίας αποκλείουν έναν άριστο και προκλητικό δείκτη αξιολόγησης για τις δικές τους τεχνικές και μεθοδολογίες, αφού σύνηθες γνώρισμα των εφαρμογών ΑΠ είναι ο πολλών διαστάσεων χώρος χαρακτηριστικών. Τα τελευταία 5 χρόνια αυτό έχει οδηγήσει όλο και περισσότερους ερευνητές ΜΜ να υιοθετούν την ΚΚ ως σημείο αξιολόγησης για την εφαρμογή της επιλογής τους, πράγμα που σημαίνει οτι οι πιο σύγχρονες τεχνικές ΜΜ έχουν εισαχθεί στην ΚΚ με ελάχιστη καθυστέρηση από την ανακάλυψή τους.

Για τους προγραμματιστές εφαρμογών, το ενδιαφέρον που υπάρχει οφείλεται κυρίως στην εξαιρετικά αυξημένη ανάγκη διαχείρισης όλο και μεγαλύτερων ποσοτήτων εγγράφων, η ανάγκη αυτή συνδέεται με την αύξηση της δικτύωσης καθώς και της μεγάλης διαθεσιμότητας εγγράφων παντός τύπου σε όλα τα επίπεδα της αλυσίδας της πληροφορίας. Όμως ενδιαφέρον υπάρχει λόγω του οτι στις τεχνικές ΚΚ έχουμε βελτίωση της απόδοσης τους, αυτά τα επίπεδα απόδοσης επιτυγχάνονται με υψηλής απόδοσης hardware/sofware πόρους. Αυτό σημαίνει οτι όλο και περισσότεροι οργανισμοί αυτοματοποιούν τις διαδικασίες που μπορούν να υλοποιηθούν με τη διαδικασία της ΚΚ.

Στο κεφάλαιο αυτό προσπαθούμε να ρίξουμε μια πιο κοντινή ματιά στην ΚΚ, περιγράφοντας την μεθοδολογία με την οποία υλοποιείται ένα σύστημα ΚΚ, καθώς να γίνει μια ανασκόπηση τεχνικών, εφαρμογών, εργαλείων και πόρων που απαιτούνται για την υλοποίηση του συστήματος.

3.2 Η βασική εικόνα

Η ΚΚ μπορεί να μορφοποιηθεί σαν τη διαδικασία προσέγγισης του στόχου της άγνωστης συνάρτησης $\Phi: D \times C \rightarrow \{T, F\}$ (που περιγράφει πως ένα κείμενο πρέπει να κατηγοριοποιηθεί σε ένα έγκυρο είδος) μέσω της συνάρτησης $\hat{\Phi}: D \times C \rightarrow \{T, F\}$ που ονομάζεται κατηγοριοποιητής,

όπου $C=\{c_{1,}\ldots c_{|C|}\}$ είναι ένα προκαθορισμένο σύνολο από κατηγορίες και D είναι ένα σύνολο κειμένων (πιθανώς με άπειρο πλήθος). Αν $\Phi(d_j,c_i)=T$, τότε το d_j αποκαλείται “θετικό παράδειγμα” (ή μέλος) του c_i , ενώ αν $\Phi(d_j,c_i)=F$ τότε ονομάζεται “αρνητικό παράδειγμα” το c_i .

Οι κατηγορίες είναι απλά συμβολικές ετικέτες: δεν είναι διαθέσιμη καμία επιπρόσθετη πληροφορία για την σημασία τους και συχνά διαθέσιμα δεν είναι ούτε τα μεταδεδομένα. Σε αυτές τις περιπτώσεις η κατηγοριοποίηση θα πρέπει να επιτευχθεί με τη βασική γνώση που μπορεί να εξαχθεί από το κείμενο αυτό καθ' αυτό. Επειδή αυτή είναι μια από τις πιο συνηθισμένες περιπτώσεις η έρευνα στο πεδίο της ΚΚ εστιάζει σε αυτή την περίπτωση και για εμάς θα αποτελέσει αντικείμενο του κεφαλαίου. Ωστόσο, όταν σε μια συγκεκριμένη εφαρμογή υπάρχουν διαθέσιμες εξωτερικές γνώσεις και μεταδεδομένα τότε μπορεί να υιοθετηθεί κάθε είδους ευρετική μέθοδος προκειμένου να έχουμε μόχλευση των δεδομένων είτε σε συνδυασμό είτε ξεχωριστά από τις τεχνικές ΑΠ και ΜΜ που συζητάμε εδώ.

Η ΚΚ είναι υποκειμενική: Όταν δύο ειδικοί αποφασίσουν ή όχι να κατηγοριοποιήσουν ένα κείμενο d_j στην κατηγορία c_i , ίσως να συμφωνήσουν, κάτι το οποίο συμβαίνει με μεγάλη πιθανότητα. Ένα δημοσιογραφικό άρθρο σχετικά με την πώληση των μετοχών του George W. Bush στην ομάδα baseball Texas Bull θα μπορούσε να ανήκει στην κατηγορία *Πολιτική* ή στην κατηγορία *Χρηματιστήριο* ή στον *Αθλητισμό*, ή σε οποιονδήποτε συνδυασμό των παραπάνω είτε σε κανένα ανάλογα με την υποκειμενική κρίση του ειδικού. Για το λόγο αυτό η σημασία μιας κατηγορίας είναι υποκειμενική, και οι τεχνικές της ΜΜ αντί να προσπαθούν να δημιουργήσουν ένα “χρυσό πρότυπο” αμφιβόλου ύπαρξης, σκοπεύουν να ανακατασκευάσουν αυτή την υποκειμενικότητα εξετάζοντας την περίπτωση, δηλαδή τα κείμενα που ο ειδικός με το χέρι κατηγοριοποίησε στην κατηγορία C. Το είδος της μάθησης αυτών των τεχνικών ΜΜ συνήθως αποκαλούνται εποπτευόμενη μάθηση, καθώς εποπτεύεται από τη γνώση των ήδη κατηγοριοποιημένων δεδομένων.

Ανάλογα την εφαρμογή, η ΚΚ μπορεί να είναι διαδικασία μιας -ετικέτας (δηλαδή ακριβώς ένα $c_i\in C$ πρέπει να ανατεθεί σε κάθε $d_j\in D$), ή πολλαπλών -ετικετών (δηλαδή για οποιοδήποτε τιμή $0\le n_j\le|C|$ κατηγοριών μπορεί να ανατεθεί σε κάθε κείμενο $d_j\in D$). Μια ειδική κατηγορία της ΚΚ με μια -ετικέτα είναι η δυαδική ΚΚ, στην οποία δίνεται μια κατηγορία c_i , κάθε $d_j\in D$ πρέπει να ανατίθεται ή στην c_i ή στην αντίθετη $\bar{c}_i$. Ο δυαδικός κατηγοριοποιητής για το c_i είναι η συνάρτηση $\hat{\Phi}_i:D\rightarrow\{T,F\}$ που προσεγγίζει τον άγνωστο στόχο της συνάρτηση $\Phi_i:D\rightarrow\{T,F\}$

Ένα πρόβλημα της ΚΚ με πολλαπλές ετικέτες με $C=\{c_{1,}\ldots,c_{|C|}\}$ συχνά αντιμετωπίζεται ως $|C|$ ανεξάρτητα προβλήματα δυαδικών κωδικοποιητών με $\{c_i,\bar{c}_i\}\gamma\iota\alpha\, i=1,\ldots,|C|$ Σε αυτή τη περίπτωση ο κατηγοριοποιητής για το C στην ουσία αποτελείται από $|C|$ δυαδικούς κατηγοριοποιητές.

Από άποψη ΜΜ, η εκπαίδευση ενός δυαδικού κωδικοποιητή είναι απλούστερη σε σχέση με την εκπαίδευση ενός κατηγοριοποιητή μιας -ετικέτας. Ως συνέπεια, όσο όλες οι κλάσεις των τεχνικών

επιβλεπόμενης ΜΜ διαχειρίζονται προβλήματα δυαδικών κατηγοριοποιητών, αφού είναι πολύ εφευρετικά, για κάποιες κατηγορίες τεχνικών μια ικανοποιητική λύση του προβλήματος μιας κλάσης συνεχίζει να είναι αντικείμενο έρευνας (Crammer and Singer, 2001).

Εκτός από πραγματική επιχειρησιακή χρήση, την οποία δεν αναλύουμε εδώ, μπορούμε περίπου να διακρίνουμε τρεις διαφορετικές φάσεις στην ζωή ενός συστήματος ΚΚ, που παραδοσιακά προσεγγίζονται απομονωμένα το ένα από το άλλο (δηλαδή η λύση του ενός προβλήματος δεν επηρεάζεται από την λύση των άλλων δύο): Ευρετήριο Εγγράφων (Document Indexing), εκπαίδευση κατηγοριοποιητή και αξιολόγηση κατηγοριοποιητή (Sebastiani, 2002).

3.3 Ευρετηριοποίηση Κειμένου (Document Indexing)

Η ευρετηριοποίηση κειμένου είναι η διαδικασία της χαρτογράφησης ενός εγγράφου d_j με μια συμπαγή αναπαράσταση του περιεχομένου του έτσι ώστε στην συνέχεια να μπορεί να ερμηνευθεί απευθείας από ένα: (1) κατηγοριοποίητή-δημιουργό αλγορίθμου, (2) ένα κατηγοριοποιητή αφού έχει δημιουργηθεί. Η ευρετερηιοποίηση κειμένου χρησιμοποιείται στην κατηγοριοποίηση κειμένου και έχει δανειστεί από την ανάκτηση πληροφορίας, όπου ένα έγγραφο d_j τυπικά αναπαριστάται με ένα διάνυσμα με βάρη: $\vec{d}_j = \langle w_{1j}, \ldots, w_{|T|j} \rangle$ όπου το Τ είναι λεξικό δηλαδή ένα σύνολο όρων (γνωστά και ως χαρακτηριστικά) που προκύπτουν τουλάχιστον μια φορά σε τουλάχιστον k κείμενα (στην κατηγοριοποίηση τουλάχιστον k κείμενα εκπαίδευσης) και το $0 \leq w_{kj} \leq 1$ ποσοτικοποιεί την σημασία του t_k που χαρακτηρίζει τη σημασιολογία του d_j . Οι τιμές που παίρνει συνήθως το k είναι μεταξύ του 1 και 5.

Μια μέθοδος ευρετηριοποίησης χαρακτηρίζεται από: (1) τον ορισμό του τι είναι ο όρος και (2) τη μέθοδο με την οποία υπολογίζονται τα βάρη των όρων. Σχετικά με το (1) συχνά επιλέγεται η αναγνώριση του όρου βάσει των λέξεων που παρουσιάζονται στο κείμενο ή με τη μορφολογική τους ρίζα που λαμβάνεται από την εφαρμογή ενός αλγορίθμου stemming (Frakes 1992). Μια δημοφιλής επιλογή είναι η προσθήκη στο σύνολο των λέξεων ενός συνόλου φράσεων, δηλαδή μεγαλύτερων γλωσσικών μονάδων που εξάγονται από το κείμενο με στατιστικές τεχνικές (Caropreso et al., 2001). Σχετικά με το (2), το βάρος των όρων μπορεί να παίρνει δυαδικές τιμές (δηλαδή $w_{kj} \in \{0,1\}$) ή μη-δυαδικές τιμές (δηλαδή $0 \leq w_{kj} \leq 1$) αναλόγως αν ο κατηγοριοποιητής-δημιουργός ή ο κατηγοριοποιητής έχουν δημιουργηθεί χρησιμοποιώντας δυαδική ή όχι είσοδο. Όταν τα βάρη είναι δυαδικά υποδεικνύουν την παρουσία ή όχι στο κείμενο. Όταν τα βάρη πέρνουν μη -δυαδικές τιμές σημαίνει οτι ο υπολογισμός τους έχει γίνει με τη χρήση μεθόδων στατιστικής ή θεωρίας πιθανοτήτων (Zobel and Moffat, 1998).

3.4 Εκπαίδευση Κατηγοριοποιητή

Ο κατηγοριοποιητής κειμένου για τη c_i δημιουργείται αυτόματα με μια γενική επαγωγική διαδικασία με την παρατήρηση των χαρακτηριστικών του συνόλου των κειμένων που είναι ηδη κατηγοριοποιημένα στη c_i ή στη $\bar{c}_i$, σταχυολογούνται τα χαρακτηριστικά τα οποία είναι απαραίτητα για ένα νέο κείμενο προκυμένου να ανήκει στη c_i . Για να κατασκευαστεί ο

κατηγοριοποιητής για το C απαιτείται ένα σύνολο Ω κειμένων τέτοια ώστε η τιμή του $\Phi(d_j, c_i)$ να είναι γνωστή για κάθε $\langle d_j, c_i \rangle \in \Omega \times C$. Στην πειραματική κατηγοριοποίηση κειμένου είναι σύνηθες το Ω να σπάει σε τρία σύνολα T_r (σύνολα εκπαίδευσης), V_α (σύνολα επιβεβαιωσης) και T_e (σύλοδα δοκιμών). Τα σύνολα εκπαίδευσης είναι το σύνολο των κειμένων με τα οποία κατασκευάζεται ο κατηγοριοποιητής. Τα δύνολα επιβεβαίωσης χρησιμοποιούνται από τον μηχανικό για μικρο-ρυθμίσεις στον κατηγοριοποιητή, δηλαδή για την εισαγωγή παραμέτρων που απαιτεί ο κατηγοριοποιητής, η αξία που αποκομίζεται προσφέρει καλύτερη αποδοτικότητα στην επαληθευση του V_α . Το σύνολο των δοκιμών είναι το σύνολο των κειμένων βάση των οποίων εξετάζεται η αποδοτικότητα του κατηγοριοποιητή.

Στην βιβλιογραφία της κατηγοριοποίησης κειμένου βλέπουμε διαφορετικού είδους κατηγοριοποιητές. Κάποιες από τις μεθόδους δημιουργούν δυαδικούς κατηγοριοποιητές της μορφής $\hat{\Phi}: D \times C \rightarrow \{T, \Phi\}$ αλλά κάποιες άλλες συναρτήσεις με πραγματικές τιμές της μορφής: $CSV: D \times C \rightarrow [0,1]$ (CSV standing for categorization status value). Για τα τελευταία, ένα σύνολο κατώτατων ορίων τ_i χρειάζεται να προσδιοριστεί για να επιτραπεί η μετατροπή των πραγματικών τιμών του CSV στις τελικές δυαδικές τιμές απόφασης (Yang, 2001).

Αξίζει τον κόπο να αναφέρουμε ότι σε αρκετές εφαρμογές μια μέθοδος υλοποιεί συνάρτηση πραγματικών τιμών, γεγονός το οποίο μπορεί να χρησιμοποιηθεί προγνωστικά.

4 Αλγόριθμοι κατηγοριοποίησης κειμενικού περιεχομένου

Στην παρούσα ενότητα αναφέρουμε τους πιο γνωστούς αλγορίθμους κατηγοριοποίησης κειμένου. Οι περίσσοτεροι διαμορφώθηκαν τη δεκαετία του ‘60, ενώ από τότε μέχρι σήμερα έχουν προκύψει πολλές τροποποιήσεις και μελέτες πάνω σε αυτούς (Χαραλαμπόπουλος, 2011). Στην παρούσα ενότητα θα ασχοληθούμε με την πρωτότυπη μορφή των αλγορίθμων.

4.1 Μπεϊσιανή Προσέγγιση (Bayesian)

Βασικό στοιχείο της συγκεκριμένης προσέγγισης είναι το Μπεϊσιανό μοντέλο. Μιλάμε για ένα μοντέλο πιθανοτήτων που συχνά τα αποτελέσματα του δεν βρίσκουν εφαρμογή στον πραγματικό κόσμο. Γενικά το πιθανοτηκό μοντέλο κατηγοριοποίησης είναι ένα μοντέλο καταστάσεων.

Για να εφαρμόσουμε το μοντέλο αυτό απαραίτητη προϋπόθεση είναι η δημιουργία μιας βάσης δεδομένων με ένα λεξικό που θα περιέχει όλες τις λέξεις που βρίσκονται στο αρχικό κατηγοριοποιημένο εκπαιδευτικό περιεχόμενο. Ας υποθέσουμε οτι τα έγγραφα θα πρέπει να κατηγοριοποιηθούν στις κατηγορίες $K_{1,} K_{2,} \ldots, K_i$. Τότε οι πιθανότητες ένα καινούργιου έγγραφό να ανήκει στην κατηγορία K_j είναι:

$$P(K_i) = \frac{\textit{πλήθος εγγράφων στην κατηγορία } K_i}{\textit{πλήθος εγγράφων}}$$

Κάθε λέξη Λ_i στο υπό μελέτη έγγραφο, τη συγκρίνουμε με το αρχικό σύνολο λέξεων

εκπαίδευσης, από τις συγκρίσεις θα έχουμε τις περιπτώσεις n_{Λ} που η λέξη ταιριάζει σε μια κατηγορία K_i . Έτσι η δεσμευμένη πιθανότητα, η λέξη Λ_j να περιέχεται στην κατηγορία K_i είναι:

$$P(\Lambda_j|K_i)=\frac{n_{\Lambda}}{n}$$

Ενώ η πιθανότητα το εκπαιδευτικό έγγραφο Ε να αντιστοιχεί στην κατηγορία K_i είναι:

$$P(E|K_i)=\prod_{j=1}^{n} P(\Lambda_j|K_i)$$

Από Μπέιες η πιθανότητα η κατηγορία K_i να περιέχει το έγγραφο Ε, γνωρίζοντας οτι ήδη έχει καταλήξει σε αυτή είναι:

$$P(K_i|E)=\frac{P(K_i)}{P(E_i)}\cdot P(E|K_i)$$

Η μέγιστη δεσμευμένη πιθανότητα θα μας δώσει και την κατηγορία στην οποία θα ανήκει και το εκπαιδευτικό έγγραφο.

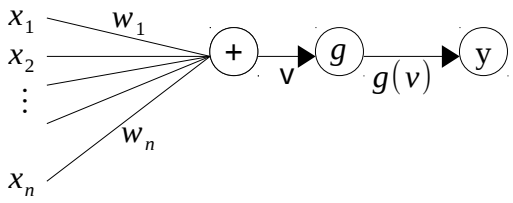
Ο συγκεκριμένος αλγόριθμος ειδικά στην κατηγοριοποίηση κειμένου παρέχει αυξημένη αποδοτικότητα σε σχέση με άλλους αλγορίθμους. Παρότι η επίδοση του είναι αντίστοιχη με αυτή των δέντρων απόφασης και των νευρωντικών δικτύων, σε περιπτώσεις κατηγοριοποίησης μεγάλων βάσεων δεδομένων παρατηρείται μεγαλύτερη ακρίβεια και ταχύτητα (Domingos and Pazzani, 1997).

4.2 Νευρωνικά Δίκτυα (Neural Networks)

Τα τεχνικά νευρωνικά δίκτυα είναι ηλεκτρονικά δίκτυα νευρώνων που έχουν ως βάση τους τη νευρωντική δομή του εγκεφάλου. Τα συγκεκριμένα δίκτυα επεξεργάζονται τις εγγραφές μεμονωμένα και κατά σειρά, ενώ μαθαίνουν συγκρίνοντας την αρχική κατηγοριοποίηση της εγγραφής, με την ορθή κατηγοριοποίηση της εγγραφής. Τα λάθη από την αρχική εκτέλεση του αλγορίθμου ανατροφοδοτούνται στο δίκτυο, έτσι ώστε να τροποποιηθούν οι υπολογισμοί για την δεύτερη εκτέλεση κ.ο.κ.

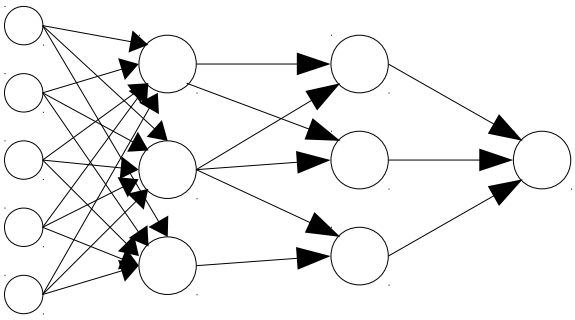
Με λίγα λόγια ένα νευρωτικό δίκτυο είναι:

1. Ένα σύνολο από τιμές εισόδου x_i με τα αντίστοιχα βάρη w_i
2. Μια συνάρτηση g που προσθέτει τα βάρη και χαρτοποιεί τα αποτελέσματα σε μια έξοδο y



Εικόνα 1: Αναπαράσταση νευρωνικου δικτύου

οι νευρώνες είναι οργανωμένοι σε επίπεδα με τον εξής τρόπο:



Εικόνα 2: Αναπαράσταση επιπέδων νευρωνικού δικτύου

Το επίπεδο εισόδου απαρτίζεται όχι από ολόκληρους νευρώνες αλλά από τις τιμές μιας εγγραφής δεδομένων, οι οποίες αποτελούν εισόδους στο επόμενο επίπεδο νευρώνων. Το επίπεδο μετά αυτό της εισόδου καλείται κρυφό επίπεδο, είναι κατ' ανάγκη μοναδικό, αλλά μπορούν να υπάρχουν παραπάνω από ένα κρυφά επίπεδα. Το τελευταίο επίπεδο είναι το επίπεδο εξόδου με ένα κόμβο για κάθε κατηγορία. Με μια ενιαία κίνηση προς τα εμπρός σαρώνονται τα αποτελέσματα του δικτύου και ανατίθενται τιμές σε κάθε κόμβο εξόδου, ώστε το έγγραφο να εκχωρείται σε όποιο κόμβο κατηγορίας έχει τη μεγαλύτερη τιμή.

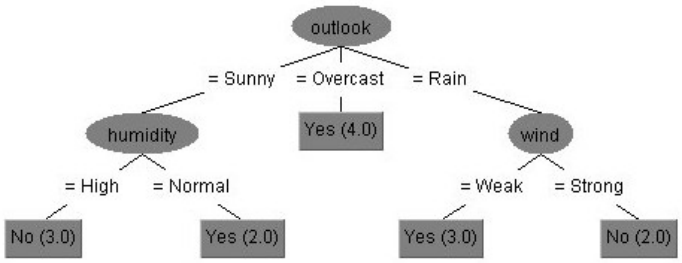
Σημαντική φάση του αλγορίθμου είναι η εκπαίδευση. Σε αυτή, η σωστή κατηγορία για κάθε έγγραφο του εκπαιδευτικού είναι πλέον γνωστή και έτσι στους κόμβους εξόδου μπορούν να ελεγχθούν οι σωστές τιμές, 1 για τον κόμβο που έχει εκχωρηθεί στη σωστή κατηγορία και 0 για τους άλλους.

Σημαντικό στοιχείο του αλγορίθμου είναι η επαναληπτική διαδικασία της μάθησης στην οποία τα δεδομένα παρουσιάζονται στο δίκτυο ένα την φορά και τα βάρη που σχετίζονται με τις τιμές εισόδου ρυθμίζονται σε κάθε εκτέλεση. Αφού μελετηθούν όλες οι σειρές η διαδικασία ξεκινάει από την αρχή.

4.3 Δέντρα Απόφασης (Decision Trees)

Μια ακόμα μέθοδος είναι η κατηγοριοποίηση κειμένου με χρήση δέντρων απόφασης. Τα δέντρα είναι εύκολα στην κατανόηση και την τροποποίηση, ενώ το ανεπτυγμένο μοντέλο μπορεί να εκφραστεί σαν ένα σύνολο κανόνων απόφασης. Σε μεγάλες βάσεις δεδομένων ο αλγόριθμος βρίσκει καλή εφαρμογή ακόμα και αν υπάρχουν διαφορετικοί αριθμοί εκπαιδευτικών δεδομένων και σημαντικός αριθμός ιδιοτήτων και χαρακτηριστικών.

Το ζητούμενο της κατηγοριοποίησης είναι να κατασκευάσουμε ένα μοντέλο της κατηγορίας αυτής με βάση όλες τις ιδιότητες. Αφού κατασκευάσουμε το μοντέλο, τότε μπορεί να χρησιμοποιηθεί για να καθορίσει την κατηγορία των εγγράφων που δεν έχουν κατηγοριοποιηθεί. Εφαρμογές της κατηγοριοποίησης με δέντρα αποφάσεων προκύπτουν σε διάφορους τομείς η λιανική πώληση, η στοχευμένες πωλήσεις και η ιατρική διάγνωση.



Εικόνα 3: Αναπαράσταση δέντρου αποφάσεων

Ένας κατηγοριοποιητής δέντρων απόφασης κατασκευάζεται με δύο βήματα, το βήμα της ανάπτυξης και το βήμα της κλάδευσης. Μετά τη δημιουργία του αρχικού δέντρου δημιουργείται ένα υπό-δέντρο με με το μικρότερο εκτιμώμενο ρυθμό σφάλματος. Η διαδικασία της κλάδευσης του αρχικού δέντρου είναι ουσιαστικά αφαιρέσεις μικρών, βαθέων κόμβων του δέντρου που οφείλονται στο θόρυβο που περιέχουν τα δεδομένα εκπαίδευσης. Έτσι μειώνεται το ρίσκο της υπέρ-συμπλήρωσης και παρέχουμε ακριβέστερη κατηγοριοποίηση των δεδομένων που έχουν ήδη κατηγοριοποιηθεί. Στο πρώτο βήμα ο στόχος είναι ο κάθε κόμβος να μπορεί να καθορίσει τις ιδιότητες διαχωρισμού και το σημείο που θα χωρίσει τα εκπαιδευτικά δεδομένα που ανήκουν σε αυτό το φύλλο. Η τιμή κάθε σημείου διαίρεσης εξαρτάται από το πόσο καλά διαχωρίζει τα δεδομένα διαφορετικών κατηγοριών. Αυτό είναι και το χαρακτηριστικό το οποίο ερευνάται συνεχώς ώστε να μπορούμε να αξιολογούμε και να αναδημιουργούμε τα σημεία διαχωρισμού. Αφού όπως μπορούμε να καταλάβουμε ο ορθός διαχωρισμός στους κόμβους θα κάνει τον αλγόριθμο αποδοτικότερο στην επίλυση των διαφόρων προβλημάτων κατηγοριοποίησης (Breiman, Leo, et al., 1984).

4.4 Αλγόριθμος Term Frequency–Inverse Document Frequency (tf-idf)

Αποτελεί το δημοφιλέστερο αλγόριθμο κατηγοριοποίησης κειμένου και χρησιμοποιείται σε πολλές εφαρμογές. Αυτός ο αλγόριθμος χρησιμοποιεί δύο βάρη, το ένα είναι η συχνότητα όρων και το άλλο η αντίστροφη συχνότητα εγγράφων. Το πρώτο βάρος δείχνει πόσο σημαντική είναι η λέξη στο έγγραφο, το οποίο βρίσκεται σε μια συλλογή εγγράφων. Η σημαντικότητα της λέξης αυξάνεται αναλογικά με τον αριθμό του πλήθους των εμφανίσεων σε ένα κείμενο, αλλά αντισταθμίζεται από την συχνότητα της εμφάνισης της λέξης στα έγγραφα της συλλογής.

Πρακτικά μπορούμε να θεωρήσουμε ένα σύνολο εγγράφων κειμένου στο οποίο θέλουμε, ποιο από όλα είναι πιο σχετικό με την αναζήτηση “όρος 1 όρος 2 όρος 3”. Ένας απλός τρόπος είναι να ξεκινήσουμε διώχνοντας όσα έγγραφα δεν περιέχουν τις τρείς λέξεις. Στην συνέχεια θα μετρήσουμε τον αριθμό εμφάνισης του κάθε όρου στο κάθε έγγραφο και θα τα αθροίσουμε όλα μαζί. Το πλήθος εμφάνισης του όρου στο έγγραφο ονομάζεται συχνότητα του όρου. Θεωρητικά κάποιος από τους τρεις όρους μπορεί να είναι ένας σύνδεσμος ή ένα άρθρο, οπότε θα έχει και μεγάλη συχνότητα στο έγγραφο μην δίνοντας έτσι χώρο σε άλλους όρους με μεγαλύτερο νόημα. Επίσης για να διαχωριστούν τα σχετικά και μη σχετικά έγγραφα δεν βοηθούν όροι που περιέχουν συνδέσμους ή άρθρα. Επομένως η αντίστροφη συχνότητα εγγράφου έχει ενσωματωθεί έτσι ώστε να αντισταθμίζει το βάρος των όρων που εμφανίζονται αρκετά συχνά στη συλλογή εγγράφων και να αυξάνει το βάρος όσων εμφανίζονται σπανιότερα.

Η αρίθμηση όρου σε ένα συγκεκριμένο κείμενο είναι όπως είπαμε ο αριθμός του πλήθους εμφανίσεων του όρου στο έγγραφο. Αυτό συνήθως κανονικοποιείται για να εμποδίσει την οποιαδήποτε κλίση προς μεγαλύτερου μήκους έγγραφα, που θα έχουν και μεγαλύτερο πλήθος όρων. Έτσι για να έχουμε ένα μέτρο της σημασίας ενός όρου o_i σε ένα έγγραφο ε_j θα ορίσουμε τη συχνότητα όρων ως εξής:

(ΣΟ) $$\textit{Συχνότητα Όρων}_{i,j} = \frac{\alpha_{i,j}}{\sum_k a_{k,j}}$$

όπου $\alpha_{i,j}$ είναι ο αριθμός των εμφανίσεων του όρου o_i στο έγγραφο ε_j και ο παρονομαστής είναι το άθροισμα όλων των όρων στο ε_j, δηλαδή το μέγεθος του κειμένου που είναι ίσο με $|\varepsilon_j|$.

Η αντίστροφη συχνότητα εγγράφου είναι ένα μέτρο που μας δείχνει τη γενική σημασία του όρου. Για τον υπολογισμό της θα διαιρέσουμε τον συνολικό αριθμό των εγγράφων, με τον αριθμό των εγγράφων που περιέχουν τον όρο και θα πάρουμε το λογάριθμο αυτού του πηλίκου, δηλαδή:

(ΑΣΕ) $$\text{Αντίστροφη Συχνότητα Εγγράφου}_{i,j} = \log \frac{|E|}{|\{\varepsilon : o_i \in \varepsilon\}|}$$

όπου $|E|$ το σύνολο των εγγράφων στη συλλογή μας και το $|\{\varepsilon : o_i \in \varepsilon\}|$ είναι ο αριθμός των εγγράφων που εμφανίζεται ο όρος o_i. Άρα τελικά ο αλγόριθμος υπολογίζει το:

$$(\text{ΣΟ} - \text{ΑΣΕ})_{i,j} = \text{ΣΟ}_{i,j} \times \text{ΑΣΕ}_{i,j}$$

όπου ένα υψηλό βάρος αυτού του όρου επιτυγχάνεται με μια υψηλή συχνότητα όρου και μια

χαμηλή συχνότητα του όρου στο σύνολο των εγγράφων, ορισμός των βαρών και ο υπολογισμός του συνολικού βάρους ενός όρου φιλτράρει τους κοινούς και αχρείαστους όρους. Η τιμή του $(\text{ΣΟ}-\text{ΑΣΕ})_{i,j}$ για έναν όρο θα είναι μεγαλύτερη του μηδενός αν και μόνο αν ο λόγος μέσα στο λογάριθμο της ΑΣΕ είναι μεγαλύτερος της μονάδας. Διάφορες μαθηματικές μορφές του αλγορίθμου μπορούν να εξαχθούν από αυτό το μοντέλο πιθανοτήτων που μιμείται την ανθρώπινη διαδικασία αποφάσεων (Spark Jones, 1972).

4.5 Μηχανές διανυσμάτων υποστήριξης (Support Vector Machine)

Οι μηχανές διανυσμάτων υποστήριξης είναι ένα σύνολο εποπτευόμενων μεθόδων μάθησης, που αναλύουν δεδομένα και αναγνωρίζουν πρότυπα. Αυτές οι μηχανές χρησιμοποιούνται ευρέως στην κατηγοριοποίηση εγγράφων. Η αρχική επινόηση του αλγορίθμου έγινε το 1995 και μετά από λίγα χρόνια προτάθηκε και η σημερινή μορφή, αυτή των απαλών ορίων. Η κλασσική μορφή του δέχεται ένα σύνολο δεδομένων και προβλέπει, για κάθε ένα, σε ποια από τις δύο διαθέσιμες κατηγορίες ανήκει.

Η διαδικασία και ο τρόπος που αυτή υλοποιείται εντάσσει τον αλγόριθμο στους δυαδικούς γραμμικούς πιθανοτικούς κατηγοριοποιητές. Όταν δοθεί στον αλγόριθμο ένα σύνολο εκπαιδευτικών δεδομένων, τότε το κάθε ένα σηματοδοτείται ως προς την κατηγορία που ανήκει (μια εκ των δύο διαθέσιμων). Έτσι ο εκπαιδευτικός αλγόριθμος κατασκευάζει ένα μοντέλο που προβλέπει το πότε η πληροφορία θα καταλήξει στην πρώτη ή στην δεύτερη κατηγορία. Διαισθητικά ένα μοντέλο του αλγορίθμου είναι μια αναπαράσταση, των δεδομένων ως σημεία του επιπέδου, σχεδιασμένα ώστε αυτά που ανήκουν σε διαφορετικές κατηγορίες (εν τέλει οι ίδιες οι κατηγορίες να είναι διαχωρισμένες) να είναι ευδιάκριτα και να χωρίζονται από ένα κενό όσο το δυνατό πλατύτερο. Τα νέα δεδομένα τοποθετούνται στον ίδιο χώρο και η κατηγορία στην οποία θα καταλήξουν εξαρτάται από τη θέση που θα έχουν στο χώρο, ουσιαστικά σε ποιά μεριά του κενού θα καταλήξουν.

Τυπικά μια μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα υπερ-επίπεδο ή ένα σύνολο υπερ-επιπέδων σε ένα μεγάλο ή άπειρων διαστάσεων χώρο που μπορεί να χρησιμοποιηθεί για ταξινόμηση. Συνεχίζοντας τη θεωρητική μελέτη του αλγορίθμου, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερ-επίπεδο που έχει την μεγαλύτερη απόσταση, από τα κοντινότερα σημεία των εκπαιδευτικών δεδομένων οποιασδήποτε κατηγορίας (γνωστό ως λειτουργικό περιθώριο), αφού γενικά όσο μεγαλύτερο το περιθώριο τόσο μικρότερο το λάθος γενίκευσης του κατηγοριοποιητή (Cortes and Vapnik, 1995).

Παρόλο που το αρχικό πρόβλημα μπορεί να οριστεί σε πεπερασμένων διαστάσεων χώρο, αυτό που συμβαίνει συνήθως είναι οτι σε πεπερασμένους χώρους τα σύνολα δεδομένων, τα οποία πρέπει να διαχωριστούν δεν είναι γραμμικά διαχωρίσιμα. Γραμμικά διαχωρίσιμα είναι δύο σημεία του επιπέδου που μπορούν να διαχωριστούν με μια ευθεία γραμμή. Γιαυτό το λόγο προτάθηκε ο αρχικός – πεπερασμένων διαστάσεων – χώρος να σχεδιαστεί σε ένα αρκετά μεγαλύτερων διαστάσεων χώρο, που θεωρητικά θα βοηθήσει το διαχωρισμό των δεδομένων και θα μας επιτρέψει να διαχωρίσουμε γραμμικά τα δεδομένα. Οι αλγόριθμοι των μηχανών διανυσμάτων υποστήριξης

χρησιμοποιούν τη σχεδίαση σε μεγαλύτερων διαστάσεων χώρους, ο κάθε ένας με τη δική του παραμετροποίηση, ώστε τα τροποποιημένα δεδομένα να μπορούν να υπολογιστούν με ευκολία. Ο υπολογισμός αυτός γίνεται σε σχέση με τα δεδομένα στον αρχικό πεπερασμένο χώρο μας, πάντα με θεωρητικό και πρακτικό όριο του διαθέσιμους υπολογιστικούς πόρους. Τα τροποποιημένα δεδομένα μας του νέου χώρου ορίζονται από μια συνάρτηση πυρήνα (kernel) $K(x,y)$, που επιλέγεται αναλόγως με τις απαιτήσεις του προβλήματος. Η συνάρτηση πυρήνα ενός συνόλου μας επιστρέφει τα ουδέτερα στοιχεία, όπως αυτά ορίζονται από τις πράξεις και το πεδίο ορισμού του συνόλου.

Τα υπερ-επίπεδα σε ένα μεγάλο χώρο ορίζονται από το σύνολο των σημείων που το γινόμενο τους είναι ένα διάνυσμα αυτού του χώρου σταθερό. Τα διανύσματα που καθορίζουν το υπερ-επίπεδο μπορούν να επιλεγούν από ένα σύνολο γραμμικών συνδυασμών με α_i παραμέτρους των εικόνων των διανυσμάτων που παρουσιάζονται στα δεδομένα μας. Με αυτή την επιλογή ενός υπερ-επιπέδου, τα σημεία x σε έναν παρεχόμενο χώρο, που είναι σχεδιασμένα πάνω στο υπερ-επίπεδο ορίζονται από τη σχέση :

$$\sum_i a_i K(x,y) = \text{σταθερό}$$

Έχει σημασία ότι, εάν η τιμή της συνάρτησης-πυρήνα γίνεται μικρότερη, όσο το y απομακρύνεται από το x, τότε κάθε στοιχείο του αθροίσματος μετράει την εγγύτητα του (υπό μέτρηση) σημείου x με το αντίστοιχο σημείο των δεδομένων μας x_i . Έτσι με βάση τη παραπάνω εξίσωση του συνόλου των πυρήνων, μπορούμε να μετρήσουμε τη σετική εγγύτητα του εξεταστέου σημείου με τα αντίστοιχα σημεία δεδομένων, τα οποία προέρχονται από το σύνολο που πρέπει να κατηγοριοποιήσουμε. Να σημειώσουμε οτι ο σχεδιασμός των σημείων x σε ένα υπερ-επίπεδο είναι υπολογιστικά απαιτητικός και σαν αποτέλεσμα αυξάνεται η πολυπλοκότητα της κατηγοριοποίησης μεταξύ των συνόλων που δεν απέχουν και βρίσκονται σχεδόν στο ίδιο επίπεδο του αρχικού χώρου.

5 Εξόρυξη Γνώσης από Κείμενα (Text Mining)

Η Εξόρυξη Γνώσης που αναλύσαμε παραπάνω έχει σαν στόχο την εξαγωγή προτύπων από δομημένα δεδομένα (βάσεις δεδομένων). Η ραγδαία διάδοση του διαδικτύου είχε σαν αποτέλεσμα με τη σειρά του την ραγδαία αύξηση της παραγωγής πληροφορίας. Έτσι προέκυψε η ανάγκη εξόρυξη γνώσης και από αδόμητα ή ημι-δομημένα κείμενα. Για την εξόρυξη γνώσης από μέσα κοινωνικής δικτύωσης, blogs, email, forums και πολλά άλλα δημιουργήθηκε ένας νέος κλάδος ο οποίος ονομάζεται Εξόρυξη Γνώσης από Κείμενα (Text Mining).

Η αρχική προσέγγιση έγινε με την κατασκευή κανόνων με το χέρι με αποτέλεσμα η δημιουργία ενός πλήρους συνόλου κανόνων να απαιτεί αρκετούς ανθρώπινους πόρους καθώς και καλή γνώση του πεδίου. Γιαυτό το λόγο στην συνέχεια έγινε χρήση Επιβλεπόμενης Μάθησης (Supervised Learning) για την δημιουργία ενός ταξινομητή. Η είσοδος του αλγορίθμου κατασκευής του ταξινομητή είναι ένα σύνολο κειμένων για κάθε κλάση.

6 Αναγνώριση Συγγραφέα (Authorship Attribution)

Η βασική ιδέα της αναγνώρισης συγγραφέα είναι οτι μετρώντας κάποια χαρακτηριστικά κειμένου, μπορούμε να διακρίνουμε κείμενα που είναι γραμμένα από διαφορετικούς συγγραφείς (Αραβαντινού, 2015). Ο τεράστιος όγκος κειμένων που είναι διαθέσιμα online, καθώς και τα κείμενα που προκύπτουν από το ηλεκτρονικό ταχυδρομείο, τα forums και τα blogs μας διευκολύνουν στην εφαρμογή της αναγνώρισης συγγραφέα. Η ποικιλία κειμένων καθώς και ο μεγάλος αριθμός συγγραφέων κάνει την εφαρμογή της αναγνώρισης συγγραφέα ακόμα πιο ενδιαφέρον υπόθεση (Stamatatos, 2009).

Μέχρι τα τέλη του '90, η έρευνα πάνω στο πεδίο επικεντρώθηκε σε προσπάθειες να καθοριστούν χαρακτηριστικά για την ποσοτικοποίηση του στυλ του γραπτού λόγου, η παραπάνω διαδικασία είναι γνωστή σαν stylometry (Holmes, 1994· Marquart et al, 2014). Έτσι προτάθηκε ένα μεγάλο σύνολο μετρικών, σε επίπεδο χαρακτήρα, λέξης, πρότασης.

Για ένα τυπικό πρόβλημα αναγνώρισης συγγραφέα θεωρούμε οτι ένα κείμενο αγνώστου συγγραφέα ανατίθεται σε ένα υποψήφιο συγγραφέα, βάσει ενός συνόλου συγγραφέων όπου για τον καθένα έχουμε διαθέσιμο κάποιο δείγμα κειμένου. Από την σκοπιά της μηχανικής μάθησης, το παραπάνω πρόβλημα μπορεί να θεωρηθεί και πρόβλημα κατηγοριοποίησης κειμένου (Sebastiani, 2002). Πέρα όμως από αυτή την προσέγγιση μπορεί να ορισθεί και ένα σύνολο ζητημάτων τα οποία είναι:

- **Επικύρωση του συγγραφέα**, δηλαδή αν ένα κείμενο έχει γραφτεί από ένα συγκεκριμένο συγγραφέα ή όχι (Koppel and Schler, 2004).
- **Ανίχνευση λογοκλοπής (Plagiarism detection)**, δηλαδή αναζήτηση των ομοιοτήτων μεταξύ δύο κειμένων (Stein and Eissen, 2007).
- **Εξαγωγή προφίλ χρήστη (User profiling)**, δηλαδή εξαγωγή δημογραφικών πληροφοριών σχετικά με το συγγραφέα, όπως ο γεωγραφικός ιδιωματισμός, η ηλικία ή το φύλο (Koppel and Argamon, 2002).

Για το πρόβλημα της αναγνώρισης συγγραφέα οι κυριότερες κατηγορίες και υποκατηγορίες χαρακτηριστικών που εξετάζονται είναι:

- **Επίπεδο χαρακτήρων**, όπως τύποι χαρακτήρων και μέθοδοι συμπίεσης.
- **Επίπεδο λέξεων**, όπως το μήκος των προτάσεων και των λέξεων και τα ορθογραφικά λάθη.
- **Συντακτικό επίπεδο**, όπως το μέρος του λόγου που ανήκει η κάθε λέξη καθώς και η δομή των προτάσεων.
- **Σημασιολογικό επίπεδο**, όπως οι σημασιολογικές εξαρτήσεις.
- **Χαρακτηριστικά βάσει της εφαρμογής**, όπως δομικά χαρακτηριστικά εξειδικευμένα που έχουν να κάνουν με το περιεχόμενο του κειμένου.

Οι διαφορετικές προσεγγίσεις μπορούν να διακριθούν σύμφωνα με το αν επεξεργάζονται κάθε κείμενο εκπαίδευσης ξεχωριστά, ή συγκεντρωτικά (ανά συγγραφέα). Συγκεκριμένα, κάποιες προσεγγίσεις ενοποιούν όλα τα διαθέσιμα κείμενα εκπαίδευσης ανά συγγραφέα σε ένα αρχείο και

στην συνέχεια εξάγουν ένα συγκεντρωτικό αποτέλεσμα αναπαράστασης στιλ γραφής για κάθε συγγραφέα. Με αυτό τον τρόπο οι διαφορές που παρουσιάζονται σε διαφορετικά κείμενα του ίδιου συγγραφέα παραβλέπονται.

Στις σύγχρονες προσεγγίσεις, η πλειοψηφία αντιμετωπίζει το κάθε κείμενο εκπαίδευσης σαν μια μονάδα η οποία συνεισφέρει ξεχωριστά στο σύστημα. Σε μια τυπική αρχιτεκτονική τέτοιου συστήματος, κάθε δείγμα κειμένου του συνόλου εκπαίδευσης αναπαριστάται με ένα διάνυσμα χαρακτηριστικών και ένα αλγόριθμος κατηγοριοποίησης εκπαιδεύεται κάνοντας χρήση του συνόλου εκπαίδευσης, ώστε να αναπτυχθεί το μοντέλο. Έτσι το μοντέλο στης συνέχεια είναι σε θέση να αναγνωρίσει τον πραγματικό συγγραφέα ενός κειμένου από έναν άγνωστο συγγραφέα. Οι συγκεκριμένοι αλγόριθμοι κατηγοριοποίησης απαιτούν πολλαπλά στιγμιότυπα εκπαίδευσης για κάθε κλάση, έτσι ώστε να εξάγουν ένα αξιόπιστο μοντέλο. Τα δείγματα του κειμένου θα πρέπει να είναι τόσο μεγάλα ώστε τα χαρακτηριστικά αναπαράστασης να εκπροσωπούν επαρκώς το στυλ τους.

Μια μέθοδος η οποία δανείζεται κάποια στοιχεία από τις δύο παραπάνω προσεγγίσεις περιγράφεται στο (Natural Language Toolkit). Ποιο συγκεκριμένα, όλα τα κείμενα εκπαίδευσης αναπαρίστανται ξεχωριστά, ωστόσο, από τα διανύσματα αναπαράστασης των κειμένων κάθε συγγραφέα υπολογίζεται ένα μέσος όρος και έτσι προκύπτει για κάθε συγγραφέα ένα μοναδικό διάνυσμα που αναπαριστά το προφίλ του.

7 Αναγνώριση Γεωγραφικού Ιδιωματισμού Συγγραφέα

Η αυτόματη αναγνώριση συγγραφέα έως τώρα εστίαζε σε χαρακτηριστικά όπως το φύλο και την ηλικία. Η μελέτες που αφορούν την αυτόματη αναγνώριση του φύλου του συγγραφέα σημειώνει και πολύ καλά αποτελέσματα στα πειράματα κατηγορίοποίησης. Τα τελευταία χρόνια η έρευνα στα προβλήματα της αυτόματης κατηγοριοποίησης συγγραφέα εστιάζει και σε άλλα χαρακτηριστικά όπως η ηλικία και η εθνικότητα.

Η ες τώρα έρευνα στο πεδίο της αναγνώρισης εθνικότητας εστιάζει στην αναγνώριση της εθνικότητας του συγγραφέα του οποίου δεύτερη γλώσσα είναι η αγγλική (Koppel et al., 2005). Σε αυτή τη διαδικασία μέσα από τον τρόπο γραφής ατόμων που δεν έχουν ως μητρική γλώσσα τα αγγλικά γίνεται προσπάθεια προσδιορισμού της χώρας προέλευσης του συγγραφέα. Το μειονέκτημα σε αυτή τη προσπάθεια είναι το ότι δεν μπορεί να αποτυπωθεί στο γραπτό λόγο η προφορά του προφορικού λόγου, κάτι το οποίο θα μπορούσε να βοηθήσει αρκετά. Σε αυτή τη διαδικασία παρατηρούμε την εισαγωγή νέων χαρακτηριστικών των οποίων δεν είχαμε δει να γίνεται χρήση τους στα έως τώρα προβλήματα αυτόματης αναγνώρισης συγγραφέα. Αυτά τα χαρακτηριστικά αποκαλούνται “Errors” και έχουν να κάνουν με λάθη που παρουσιάζονται στον γραπτό λόγο ανθρώπων που δεν έχουν ως μητρική γλώσσα τα αγγλικά. Τα λάθη τα οποία εξετάζονται χωρίζονται στις εξής κατηγορίες:

- **Ορθογραφικά** – ορθογραφικά λάθη όπως:
    - επανάληψη του ίδιου γράμματος (δηλαδή *remmit* αντί για *remit*)
    - διπλά γράμματα που γράφονται μόνο μια (δηλαδή *comit* αντί για *commit*)

- χρήση του γράμματος α αντί για το β (δηλαδή *firsd* αντί για *first*)
- αναστροφή γραμμάτων (δηλαδή *fisrt* αντί για *first*)
- γράμματα που παρεμβάλονται (δηλαδή *friegnd* αντί για *friend*)
- γράμματα που λείπουν (δηλαδή *frend* αντί για *friend*)
- ενωμένες λέξεις (δηλαδή *stucktogether*)

- **Συντακτικά**
  - επανάληψη λέξεις
  - σύγχυση *that/which*
  - λέξεις που λείπουν
  - εσφαλμένη χρήση ενικού/πληθυντικού
  - εσφαλμένη χρήση χρονικών τύπων

8 Επιλογή χαρακτηριστικών

Το πρόβλημα της επιλογής χαρακτηριστικών από μια άποψη μπορεί να χαρακτηριστεί και πρόβλημα βελτιστοποίησης αφού αναζητείται το βέλτιστο υποσύνολο χαρακτηριστικών σε σχέση με κάποιο κριτήριο αξιολόγησης, όπως η απόδοση ενός ταξινομητή ή μια άλλη συνάρτηση αξιολόγησης. Αναλυτικότερα το πρόβλημα μπορεί να οριστεί ως εξής:

Δεδομένου ενός συνόλου διανυσμάτων $X_d=\{x_i|i=1\ldots d\}$ επέλεξε εκείνο το υποσύνολο $Y_m=\{x_{i1},x_{i2},\ldots,x_{i\mathrm{m}}\}$ με m<<d το οποίο βελτιστοποιεί μια συνάρτηση αξιολόγησης J.

$$\begin{bmatrix}x_1\\x_2\\\vdots\\x_d\end{bmatrix}\rightarrow\begin{bmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{i\mathrm{m}}\end{bmatrix}\{x_{i1,}x_{i2,}\ldots x_{i\mathrm{m}}\}=\underset{m,J_m}{\arg\max}\left[J\{x_i|i=1\ldots d\}\right]$$

Όλα τα δυνατά υποσύνολα χαρακτηριστικών ορίζουν τον χώρο των υποψήφιων λύσεων. Με δεδομένο ότι ο αριθμός των υποσυνόλων αυτών αυξάνεται εκθετικά με το πλήθος των χαρακτηριστικών N και γίνεται ίσος με 2^N κάνει το πρόβλημα της αποτίμησης όλων αυτών των συνόλων πρακτικά ασύμφορο και σε μερικές περιπτώσεις αδύνατον να λυθεί. Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιούνται διάφορες τεχνικές αναζήτησης που βασίζονται σε διάφορους αλγόριθμους αναζήτησης ντετερμινιστικούς ή στοχαστικούς αντίστοιχα.

Η επιλογή ενός βέλτιστου υποσυνόλου χαρακτηριστικών απαιτεί δύο βασικά στοιχεία που πρέπει να καθοριστούν

- τον τρόπο με τον οποίο θα γίνει η αναζήτηση των υποψήφιων λύσεων
- την αντικειμενική συνάρτηση αξιολόγησης βάση της οποίας γίνεται η επιλογή των βέλτιστων υποσυνόλων

Όσο αφορά τον τρόπο αναζήτησης των υποψήφιων λύσεων υπάρχουν οι ακόλουθες 3 στρατηγικές αναζήτησης:

- Exhaustive search (εξαντλητική αναζήτηση)
- Heuristic (Ευρετικές)
- Randomized (στοχαστικές)

Όσο αφορά τη μεθοδολογία αξιολόγησης για την επιλογή των βέλτιστων υποσυνόλων των χαρακτηριστικών οι μέθοδοι επιλογής χωρίζονται σε τρεις κατηγορίες:

- Filter
- Wrapper
- Embedded

8.1 Αλγόριθμος RELIEF-F

Ο αλγόριθμος RELIEF-F αποτελεί την εξέλιξη του αλγορίθμου RELIEF που εφάρμοσαν αρχικά οι Kira και Rendell. Ο αλγόριθμος RELIEF-F ανήκει στην κατηγορία των φίλτρων εξετάζοντας συσχετίσεις των τιμών των μεταβλητών ανάμεσα σε διαφορετικά στιγμιότυπα του προβλήματος. Μπορεί να διαχειριστεί και περιπτώσεις multi-class προβλημάτων επιλογής χαρακτηριστικών(Robnik-Šikonja et al, 2003)

Τα βήματα του αλγορίθμου περιγράφονται ως εξής:

```
1  set all weights W[A] := 0.0;
2  for i := 1 to m do begin
3      randomly select an instance Ri ;
4      find k nearest hits Hj ;
5      for each class C≠class(Ri) do
6          from class C find  k nearest misses  Mj(C) ;
7  for A:= 1 to a do
8  W[A] := W[A] − Σ_{j=1}^{k} diff(A,Ri,Hj)/(m·k) +
   Σ_{C≠class(Ri)} [P(C)/(1−P(class(Ri))) Σ_{j=1}^{k} diff(A,Ri,Mj(C))]/(m·k) ;
9  end;
```

Πίνακας 1: Ψευδοκώδικας του βασικού αλγορίθμου Relief

Η κύρια ιδέα του αλγορίθμου Relief εντοπίζεται στον υπολογισμό της διαχωριστικής ικανότητας των χαρακτηριστικών βάσει των τιμών τους σε κοντινά στιγμιότυπα του συνόλου εκπαίδευσης. Το σκεπτικό είναι το εξής: ένα χαρακτηριστικό με μεγάλη διακριτική ικανότητα, πρέπει να λαμβάνει διαφορετικές τιμές μεταξύ παραδειγμάτων που ανήκουν σε διαφορετικές κατηγορίες και να έχει την ίδια τιμή για παραδείγματα της ίδιας κατηγορίας.

Για ένα τυχαία επιλεγμένο στιγμιότυπο R_i , η μέθοδος ψάχνει τους k κοντινότερους γείτονές του (k-nearest neighbors) από τα στοιχεία του συνόλου εκπαίδευσης ως εξής:

- k παραδείγματα του συνόλου εκπαίδευσης που είναι κοντά στο Ri και ανήκουν στην ίδια κατηγορία με αυτό. Τα k αυτά παραδείγματα ονομάζονται κοντινότερες επιτυχίες (nearest hits) και,
- k παραδείγματα από διαφορετική κατηγορία τα οποία βρίσκονται σε ελάχιστη απόσταση από το Ri και ονομάζονται κοντινότερες αποτυχίες (nearest miss).

Τα m και k είναι παράμετροι που καθορίζονται από το χρήστη. Με m συμβολίζουμε το πόσες φορές θα αντλήσουμε παραδείγματα από ένα σύνολο εκπαίδευσης (ένα παράδειγμα κάθε φορά) και με k τον αριθμό των κοντινότερων γειτόνων του επιλεγμένου παραδείγματος Ri . Με a συμβολίζουμε τον αριθμό των χαρακτηριστικών.

Η κύρια φιλοσοφία του αλγορίθμου συμπυκνώνεται στις γραμμές 8-9 και είναι η εξής: Αν το παράδειγμα R_i και τα k κοντινότερα παραδείγματα $H_j \in class(R_i)$ μοιράζονται διαφορετικές τιμές για ένα χαρακτηριστικό A τότε το χαρακτηριστικό A διαχωρίζει παραδείγματα που ανήκουν στην ίδια κλάση και αυτό δεν είναι επιθυμητό από τον αλγόριθμο και έτσι μειώνεται η ποσότητα $W[A]$.

Αντίθετα αν το παράδειγμα R_i και τα τα k κοντινότερα παραδείγματα $M_j(C)$, όπου $C \notin class(R_i)$ μοιράζονται διαφορετικές τιμές για ένα χαρακτηριστικό A τότε το χαρακτηριστικό A διαχωρίζει παραδείγματα που ανήκουν σε διαφορετικές κατηγορίες και αυτό είναι επιθυμητό από τον αλγόριθμο και έτσι μεταβάλλεται θετικά η ποσότητα W[A]. Επομένως, μία μεγάλη τιμή $W[A]$ σηματοδοτεί ένα χαρακτηριστικό με καλή διακριτική ικανότητα.

Η συνάρτηση diff ορίζει τη διαφορά στις τιμές ενός χαρακτηριστικού A ανάμεσα σε δύο παραδείγματα I_1 και I_2 . Αναλυτικότερα:

Για ονομαστικές μεταβλητές έχει τη μορφή:

$$diff(A, I_{1,} I_2) = \begin{cases} 0\,; value(A, I_1) = value(A, I_2) \\ 1\,; διαφορετικά \end{cases}$$

Για αριθμητικές μεταβλητές έχει τη μορφή:

$$diff(A, I_{1,} I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)}$$

Επίσης η συνάρτηση diff χρησιμοποιείται επίσης για την εύρεση των k κοντινότερων παραδειγμάτων του επιλεγμένου παραδείγματος R_i . Η συνολική απόσταση ενός παραδείγματος I από το R_i υπολογίζεται ως το άθροισμα των αποστάσεων για όλες τα χαρακτηριστικά (μεταβλητές).

$$dist(R_i, I) = \sum_{A=1}^{\alpha} diff(A, R_i, I)$$

Κεφάλαιο 3 - Μεθοδολογία

1 Μεθοδολογία

Για την πραγματοποίηση της παρούσας διπλωματικής κάναμε μια αρκετά εξαντλητική βιβλιογραφική μελέτη. Μετά από αναζήτηση επιστημονικών άρθρων αλλά και συγγραμμάτων πραγματοποιήθηκε μελέτη στα πεδία της εξόρυξης γνώσης, της κατηγοριοποίησης κειμένου και αυτόματης αναγνώρισης συγγραφέα.

Στην συνέχεια προχωρήσαμε στην συλλογή των απαραίτητων δεδομένων, δηλαδή κειμένων από μέσα κοινωνικής δικτύωσης (facebook, twitter). Η προ επεξεργασία περιελάμβανε την απομόνωση του καθαρού κειμένου κάθε φορά. Η επισημείωση του κάθε κειμένου ήταν μια διαδικασία κατά την οποία στο κάθε κείμενο επισημάνθηκαν διάφορες ιδιότητες σχετικές με το συγγραφέα όπως φύλο, ηλικία, κ.ά Μετά την συλλογή των κειμένων και με βάση τη βιβλιογραφική μελέτη προχώρησα στον καθορισμό των χαρακτηριστικών για τον υπολογισμό των οποίων έγινε χρήση των βιβλιοθηκών του NLTK[1] και συνεπώς και η υλοποίηση έγινε με χρήση της γλώσσας προγραμματισμού Python.

Μετά την εκτέλεση των πειραμάτων δημιουργήθηκε ένα αρχείο με τα αποτελέσματα το οποίο στην συνέχεια με την βοήθεια του εργαλείου WEKA[2] προχωρήσαμε στη διαδικασία της κατηγοριοποίησης. Στα πειράματα της κατηγοριοποίησης χρησιμοποιήθηκε πληθώρα αλγορίθμων, στη συνέχεια συγκεντρώθηκαν τα αποτελέσματα της κατηγοριοποίησης και τέλος έγινε σύγκριση ως προς την αποδοτικότητα των αλγορίθμων αναδεικνύοντας τον καλύτερο.

2 Η γλώσσα προγραμματισμού Python

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού η οποία δημιουργήθηκε από τον Guido van Rossum το 1990. Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της και το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ότι θα ήταν δυνατόν σε γλώσσες όπως η C++ ή η Java. Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες και για την ταχύτητα εκμάθησης της.

Οι διερμηνείς της Python είναι διαθέσιμοι για εγκατάσταση σε πολλά λειτουργικά συστήματα, επιτρέποντας στην Python την εκτέλεση κώδικα σε ευρεία γκάμα συστημάτων. Χρησιμοποιώντας εργαλεία τρίτων, όπως το Py2exe ή το Pyinstaller, ο κώδικας της Python μπορεί να πακεταριστεί σε αυτόνομα εκτελέσιμα προγράμματα για μερικά από τα πιο δημοφιλή λειτουργικά συστήματα, επιτρέποντας τη διανομή του βασισμένου σε Python λογισμικού για χρήση σε αυτά τα περιβάλλοντα χωρίς να απαιτείται εγκατάσταση του διερμηνέα της Python.

Η Python αναπτύσσεται ως ανοιχτό λογισμικό (open source) και η διαχείρισή της γίνεται από τον

1 http://www.nltk.org/
2 http://www.cs.waikato.ac.nz/ml/weka/

μη κερδοσκοπικό οργανισμό Python Software Foundation[3]. Ο κώδικας διανέμεται με την άδεια Python Software Foundation License η οποία είναι συμβατή με την GPL. Το όνομα της γλώσσας προέρχεται από την ομάδα άγγλων κωμικών Monty Python[4].

2.1 Δομή και Σύνταξη

Η γλώσσα χρησιμοποιεί μεταγλωττιστή (compiler) για την δημιουργία του εκτελέσιμου κώδικα και σχετίζεται με τις γλώσσες προγραμματισμού Tcl, Perl, Scheme, Java και Ruby, καθώς και με την ABC η οποία υπήρξε η αρχική πηγή έμπνευσης για τη δημιουργία της.

Ένα ιδιαίτερο χαρακτηριστικό της γλώσσας είναι η χρήση κενών διαστημάτων (whitespace) για τον διαχωρισμό των συντακτικών δομών που προγράμματος, σε αντίθεση με την πρακτική σε άλλες γλώσσες όπου για τον ίδιο σκοπό χρησιμοποιούνται ειδικά σύμβολα (πχ αγκύλες). Αυτό, σε συνδυασμό με το ότι χρησιμοποιεί πλήρεις αγγλικές λέξεις στη θέση συμβόλων, καθιστούν τον κώδικα της Python ευανάγνωστο από όσους έχουν βασική γνώση των αγγλικών.

3 Εργαλεία υλοποίησης

3.1 NLTK - Natural Language Toolkit

Το NLTK αποτελεί μια πλατφόρμα γραμμένη σε Python που έχει ένα σύνολο βιβλιοθηκών για συμβολική και στατιστική επεξεργασία φυσικού κειμένου. Συγγραφείς είναι οι Steven Bird, Edward Loper, Ewan Klein και η πρώτη έκδοση βγήκε το 2001. Περιλαμβάνει έτοιμα δείγματα δεδομένων καθώς και την δυνατότητα γραφικών απεικονίσεων. Το NLTK προορίζεται για την υποστήριξη της έρευνας στην επεξεργασία φυσικής γλώσσας και σε συναφείς τομείς όπως ανάκτηση πληροφορίας, μηχανικής μάθησης και τεχνητής νοημοσύνης.

Για τις ανάγκες της υλοποίησης χρησιμοποιήσαμε τα εξής εργαλεία:

- Χωρισμός κειμένου σε λέξεις (Word Tokenization)
- Υπολογισμός εμφάνισης λέξεων μέσα στο κείμενο

3.2 Weka - Waikato Environment for Knowledge Analysis

Το Weka είναι μια σουίτα, η οποία περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και την προγνωστική μοντελοποίηση, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες. Η αρχική μη-Java έκδοση του Weka ήταν ένα Tcl/Tk front-end (ως επί το πλείστον τρίτων) για μοντελοποίηση αλγορίθμων που εφαρμόζονται σε άλλες γλώσσες προγραμματισμού, περιέχοντας δυνατότητες προεπεξεργασίας δεδομένων σε C, και ένα σύστημα βασισμένο σε Makefile για τη πραγματοποίηση πειραμάτων μηχανικής μάθησης. Αυτή η αρχική έκδοση είχε σχεδιαστεί ως ένα εργαλείο για την ανάλυση των δεδομένων από γεωργικούς τομείς, αλλά η πιο πρόσφατη πλήρης έκδοση βασισμένη σε Java (Weka 3), η ανάπτυξη της οποίας άρχισε το 1997, έχει πλέον πολλούς τομείς εφαρμογής,κυρίως

3 https://www.python.org/psf/
4 https://www.youtube.com/watch?v=jHPOzQzk9Qo

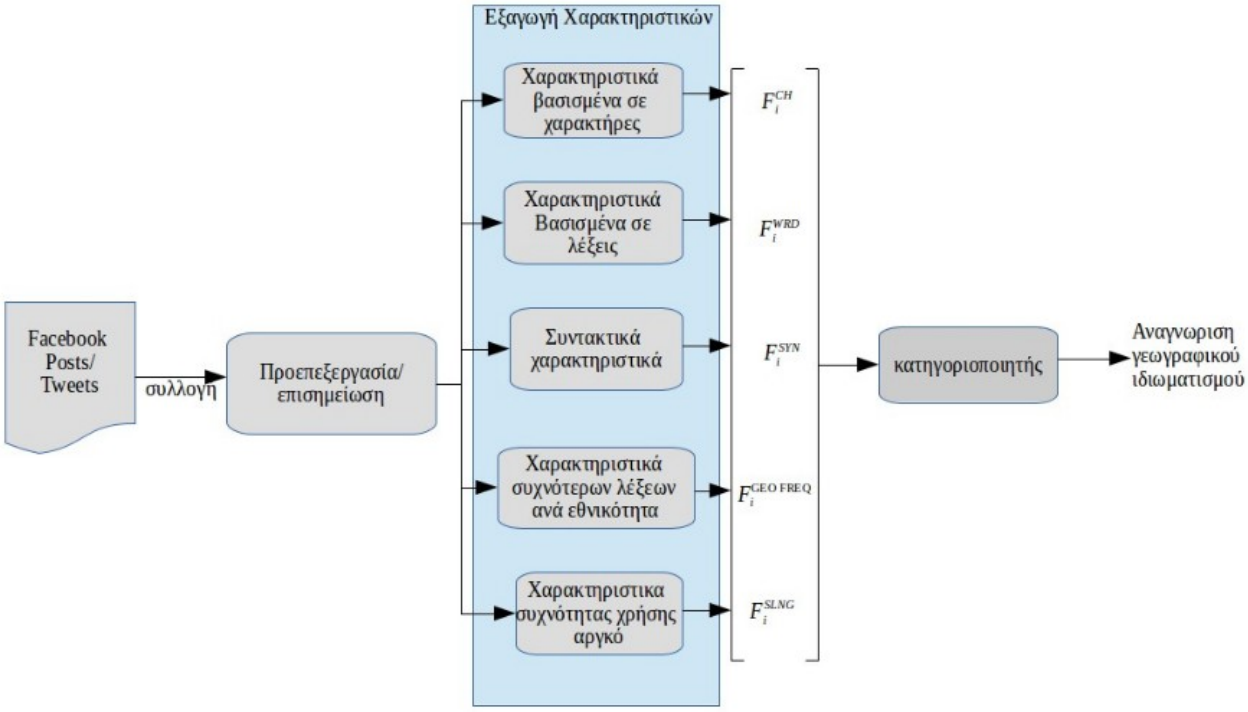
εκπαιδευτικούς σκοπούς και έρευνες. Στα πλεονεκτήματα του Weka περιλαμβάνονται:

- Δωρεάν διαθεσιμότητα υπό την GNU Γενική Άδεια Δημόσιας χρήσης.
- Φορητότητα, δεδομένου ότι έχει υλοποιηθεί πλήρως στην γλώσσα προγραμματισμού Java και έτσι τρέχει σε σχεδόν κάθε σύγχρονη υπολογιστική πλατφόρμα.
- Μια ολοκληρωμένη συλλογή δεδομένων προεπεξεργασίας και τεχνικές μοντελοποίησης.
- Ευκολία στη χρήση λόγω των γραφικών διεπαφών χρήστη.

Το Weka υποστηρίζει διάφορες βασικές διεργασίες εξόρυξης δεδομένων· πιο συγκεκριμένα, προεπεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, απεικόνιση, και την δυνατότητα επιλογής. Όλες οι τεχνικές του Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα απλό αρχείο ή συσχέτιση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό των χαρακτηριστικών (κανονικά, αριθμητικά ή ονομαστικά χαρακτηριστικά, αλλά και κάποιοι άλλοι τύποι χαρακτηριστικών υποστηρίζονται επίσης).Το Weka παρέχει πρόσβαση σε SQL βάσεις δεδομένων , χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφονται από ένα ερώτημα βάσης δεδομένων. Δεν είναι ικανό για εξόρυξη από πολυ-σχεσιακές βάσεις δεδομένων, αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής συνδεδεμένων πινάκων της βάσης δεδομένων σε έναν πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka. Άλλη μια σημαντική περιοχή που δεν καλύπτεται προς το παρόν από τους αλγορίθμους που περιλαμβάνονται στο Weka είναι η μοντελοποίηση αλληλουχιών.

Κεφάλαιο 4 – Πειραματική Διαδικασία

Στο προηγούμενο κεφάλαιο έγινε συνοπτική περιγραφή του τρόπου που δουλέψαμε για την περάτωση της παρούσας εργασίας. Σε αυτό το κεφάλαιο γίνεται αναλυτική περιγραφή της πειραματικής διαδικασίας. Μετά το πέρας μπορούμε να πούμε ότι η μεθοδολογία μπορεί να πάρει τη μορφή συστήματος του οποίου μια γενική περιγραφή φαίνεται στο παρακάτω σχήμα.



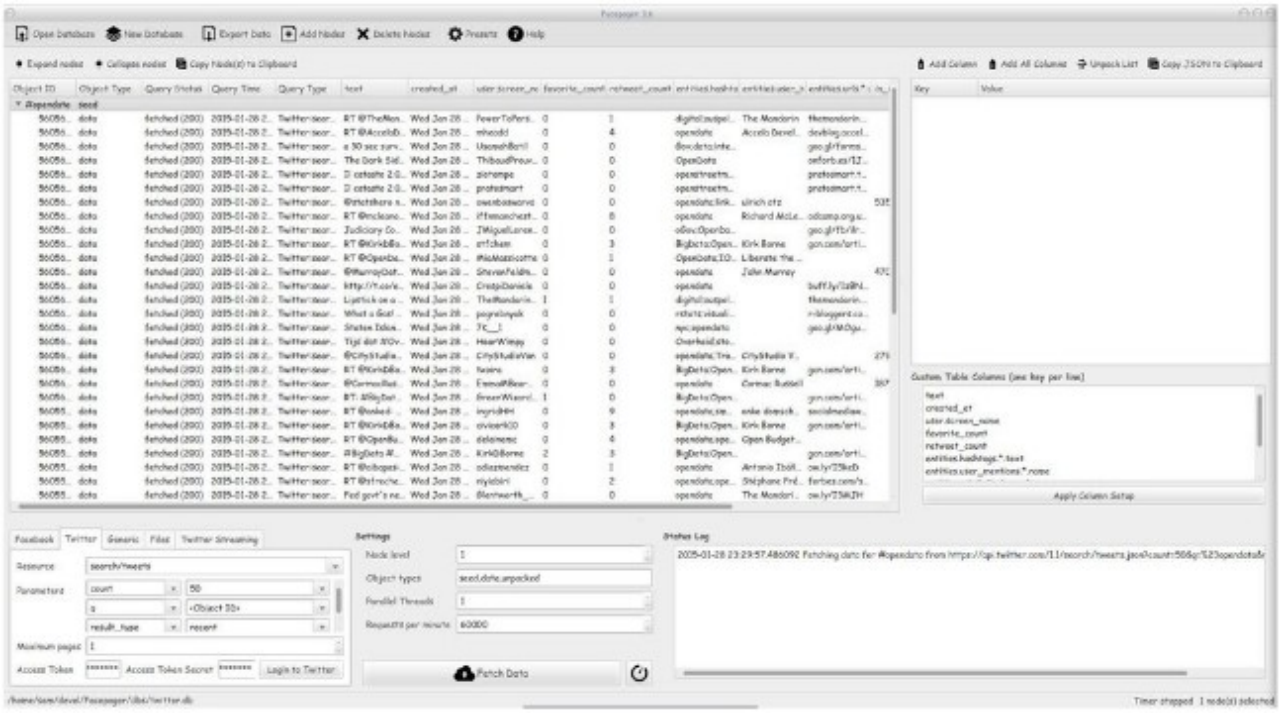
Εικόνα 4: Διάγραμμα μεθοδολογίας αναγνώρισης ως σύστημα

1 Συλλογή Δεδομένων

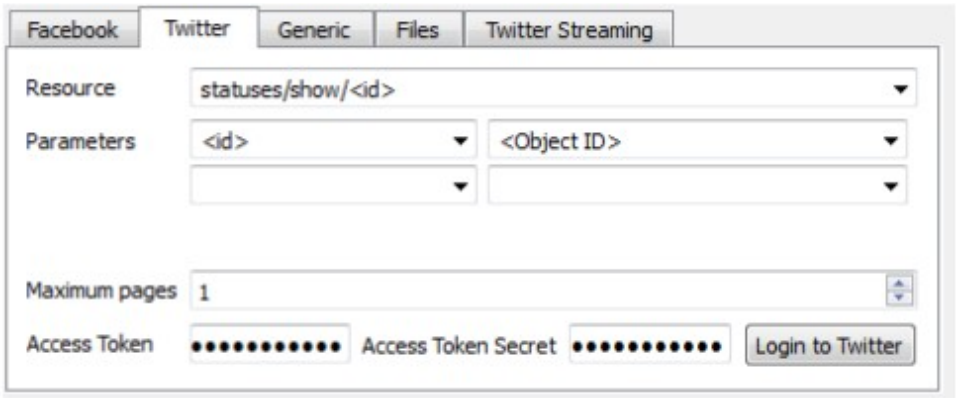
Το πρώτο στάδιο της πειραματικής διαδικασίας είναι η συλλογή δεδομένων. Στην περίπτωση μας τα δεδομένα που θέλουμε να επεξεργαστούμε είναι καθαρό κείμενο. Με τον όρο καθαρό κείμενο εννοούμε οτι η μόνη πληροφορία που φέρουν τα δεδομένα είναι το κείμενο αυτό καθαυτό. Οι πηγές από τις οποίες αλιεύθηκαν τα κείμενα ήταν από ιστότοπους κοινωνικής δικτύωσης (facebook, twitter). Τα κείμενα προέρχονται σε μεγάλο βαθμό από official accounts ατόμων για τα οποία υπάρχουν αρκετές πληροφορίες. Καλλιτέχνες, μουσικοί, πολιτικοί, ακτιβιστές είναι λίγες από τις κατηγορίες που επιλέχθηκαν για την συλλογή των δεδομένων. Προϋπόθεση για την επιλογή του συγγραφέα ήταν να υπάρχουν πληροφορίες φύλου, ηλικίας, ιδιότητας και γεωγραφικού ιδιωματισμού.

Το σύνολο των δεδομένων είναι 252.112 διαφορετικά κείμενα τα οποία στο σύνολο τους περιέχουν 34.969.115 χαρακτήρες και 5.372.512 λέξεις. Ο αριθμός των διαφορετικών συγγραφέων είναι 357. Για την συλλογή χρησιμοποιήθηκε το εργαλείο με όνομα *Facepager.* Το συγκεκριμένο εργαλείο

έχει σχεδιαστεί για να “τραβάει” δημόσια δεδομένα από τα API των Facebook, Twitter και άλλα JASON-Based API. Το εργαλείο διαθέτει γραφικό περιβάλλον στο οποίο εισάγεται με link ο λογαριασμός από το οποίο θα γίνει η συλλογή δεδομένων και στην συνέχεια βάσει των παραμέτρων που θα εισάγονται ξεκινάει η λήψη των δεδομένων. Για την αποθήκευση των δεδομένων το εργαλείο δημιουργεί τοπικές SQLite βάσεις δεδομένες.



Εικόνα 5: Facepager: GUI



Εικόνα 6: Facepager: Πεδία παραμέτρων

Στην συνέχεια πραγματοποιήθηκε η προ-επεξεργασία και η επισημείωση των κειμένων. Κατά την προ επεξεργασία έπρεπε από τη βάση δεδομένων που δημιουργήθηκε από το *facepager* να

απομονωθεί το καθαρό κείμενο τού κάθε post ή comment, καθώς η βάση δεδομένων για κάθε εγγραφή περιείχε μια πληθώρα πληροφοριών και μετα-δεδομένων που για το σκοπό της παρούσας διπλωματικής δεν έχουν σημασία.

Στην συνέχεια έγινε η επισημείωση του κειμένου. Αρχικά αξίζει να περιγραφεί η γενική μορφή του τελικού επισημειωμένου αρχείου:

Index κειμένου	Καθαρό κείμενο	Επισημειωμένα χαρακτηριστικά
1		
⋮		

Πίνακας 2: Γενική μορφή του επισημειωμένου αρχείου

Με βάση τη βιβλιογραφική μελέτη για την επισημείωση των κειμένων τα χαρακτηριστικά που επιλέχθηκαν είναι τα εξής:

1. **Φύλο Συγγραφέα**

   “*F*” για άντρας, “*M*” για γυναίκα

2. **Κατηγορία ηλικίας συγγραφέα**

   Για να έχω καλύτερα αποτελέσματα επέλεξα να χωρίσω τις ηλικίες στις εξής ομάδες:

Ηλικιακή Κατηγορία	**Εύρος Ηλικιών**
A	14 - 19
B	20 - 24
C	25 - 34
D	35 - 44
E	45 - 59
F	> 60

Πίνακας 3: Κατηγορίες Ηλικιών

3. **Ακριβής Ηλικία συγγραφέα**
4. **Κοινωνικό δίκτυο** από το οποίο αλιεύθηκε το κείμενο. Στο συγκεκριμένο σύνολο κειμένων όπως προαναφέρθηκε τα κείμενα προέρχονται από το Facebook και το Twitter.
5. **Θεματική Περιοχή**: η θεματική περιοχή περιγράφει την ιδιότητα του συγγραφέα καθώς και τον χώρο στον οποίο κινείται. Η τιμές που παίρνει το συγκεκριμένο πεδίο είναι οι κατηγορίες των pages στο Facebook δηλαδή:
   - ***Actor/Director***
   - ***Artist***

- ***Athlete***
- ***Author***
- ***Business Person***
- ***Chef***
- ***Coach***
- ***Doctor***
- ***Entertainer***
- ***Journalist***
- ***Lawyer***
- ***Musician/Band***
- ***Politician***
- ***Teacher***
- ***Writer***

6. **Εθνικότητα/ Γεωγραφικός ιδιωματισμός**: Το πεδίο αυτό περιλαμβάνει την εθνικότητα ή την μητρική γλώσσα του συγγραφέα. Επειδή όλη η υλοποίηση της διπλωματικής γίνεται για αγγλικό κείμενο σε αυτό το πεδίο έχουμε τις κυριότερες εθνικότητες με επίσημη γλώσσα την αγγλική. Δηλαδή το παρόν πεδίο παίρνει τις ετικέτες:

   "***US***" για Αμερική

   "***CAN***" για Καναδά

   "***UK***" για Αγγλία

   "***AUS***" για Αυστραλία

για συγγραφείς που δεν κατάγονται από κάποια από τις παραπάνω χώρες ή απλά δεν έχουν σαν μητρική γλώσσα την αγγλική χρησιμοποιούμε την ετικέτα:

"**NNS**" (Non native speakers).

7. **Επιπλέον Πληροφορίες**: Σε αυτό προσθέτουμε οτι επιπρόσθετες πληροφορίες έχουμε για τον συγγραφέα σχετικά με την ιδιότητα του ή το μορφωτικό του επίπεδο.

Αξίζει να σημειωθεί ότι αρκετά από τα χαρακτηριστικά που επισημειώθηκαν δεν έχουν άμεσα σχέση με την πειραματική διαδικασία όμως επέλεξα να τα εισάγω προκειμένου το ίδιο σύνολο δεδομένων στο μέλλον να είναι διαθέσιμο για έρευνα και σε άλλα πεδία.

Για την μορφοποίηση και αρχειοθέτηση του συνόλου δεδομένων που περιγράφεται παραπάνω

χρησιμοποιήθηκε αρχείο της μορφής .csv (Comma-separated Values), καθώς αποτελεί ένα απλού τύπου αρχείο δεδομένων το οποίο καλύπτει τις ανάγκες της υλοποίησης και είναι και εξαιρετικά απλός της επεξεργασία του. Σε αυτό τον τύπο δεδομένων έχουμε τη γενική μορφή ενός υπολογιστικού φύλλου του οποίου τα στοιχεία της κάθε στήλης χωρίζονται με κόμμα.

Επομένως βάσει της περιγραφής της διαδικασίας της προ-επεξεργασίας καθώς και της διαδικασίας της επισημείωσης το τελικό αρχείο με το σύνολο δεδομένων θα είναι σε μορφή υπολογιστικού φύλλου και συγκεκριμένα τα πεδία είναι της μορφής:

Index εγγράφου 1-260K	**Καθαρό κείμενο**	**Φύλο συγγραφέα** (F/M)	**Ηλικιακή κατηγορία συγγραφέα** (A/B/C/D/E/F)	**Ακριβής ηλικία συγγραφέα** >14	**Κοινωνικό δίκτυο** (Facebook/ Twitter)	**Θεματική περιοχή**	**Εθνικότητα** (US/CAN/ UK/AUS/NS)	**Επιπλέον πληροφορίες**
1								
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
263000								

Πίνακας 4: Ακριβής μορφή επισημειωμένου αρχείου

Ενώ το ακριβές πρότυπο του αρχείου είναι:

Εικόνα 7: Ακριβές πρότυπο επισημειωμένου αρχείου

2 Χαρακτηριστικά Εγγράφου

Σε αυτή την ενότητα θα γίνει ανάλυση της λογικής με βάση την οποία έγινε η επιλογή των χαρακτηριστικών του κειμένου αλλά και η διαδικασία με την οποία εξάγαμε τα χαρακτηριστικά του καθαρού κειμένου. Σε αυτή την ενότητα γυρνάμε στις βασικές αρχές της εξόρυξης γνώσης από κείμενο και προς το παρόν “ξεχνάμε” την διαδικασία της επισημείωσης που πραγματοποιήσαμε παραπάνω καθώς κατά την παρούσα διαδικασία την πληροφορία καλούμαστε να την λάβουμε από το καθαρό κείμενο αποκλειστικά. Σε μια γενική προσέγγιση του προβλήματος της αυτόματης αναγνώρισης συγγραφέα με βάση την ηλικία η επιλογή των χαρακτηριστικών θα ήταν σαφώς ευκολότερη καθώς υπάρχει και αρκετό ερευνητικό υλικό.

2.1 Εξαγωγή Χαρακτηριστικών

Η εξαγωγή των χαρακτηριστικών (feature extraction) ξεκινάει από κάποια βασικά χαρακτηριστικά τα οποία έχουν επιλεγεί από μια σειρά ερευνών του πεδίου και δείχνουν οτι συμβάλουν σε μεγάλο βαθμό στα αποτελέσματα της κατηγοριοποίησης. Τα είδη των χαρακτηριστικών που επιλέχτηκαν για την υλοποίηση μπορούν να χωριστούν σε τρεις κατηγορίες τις οποίες αναλύουμε στον

παρακάτω πίνακα.

- **Λεξιλογικά Χαρακτηριστικά:** Τα συγκεκριμένα χαρακτηριστικά μπορούν με την σειρά του ς να χωριστούν στα χαρακτηριστά βασισμένα στους χαρακτήρες και στα χαρακτηριστικά βασισμένα στις λέξεις.
- **Συντακτικά Χαρακτηριστικά:** Περιλαμβάνει τις λειτουργικές λέξεις, τα σημεία στίξης, τα μέρη του λόγου. Με τα συγκεκριμένα μπορεί να αποτυπωθεί ο τρόπος γραφής του συγγραφέα σε επίπεδο πρότασης. Το στοιχείο αυτών των χαρακτηριστικών είναι ότι διαφέρουν ανάλογα με τις συνήθειες του συγγραφέα στο τρόπο που δομεί τις προτάσεις του. Στην παρούσα διπλωματική δεν θα χρησιμοποιήσουμε μέρη του λόγου για τη διαμόρφωση των χαρακτηριστικών.
- **Δομικά Χαρακτηριστικά:** τα χαρακτηριστικά αντιπροσωπεύουν τον τρόπο που ο συγγραφέας οργανώνει την διάταξη του γραπτού του.
- **Χαρακτηριστικά με βάση το περιεχόμενο:** Για τις ανάγκες της διπλωματικής χρειάζονται χαρακτηριστικά που εστιάζουν στις διαφοροποιήσεις του γραπτού λόγου σε σχέση με την εθνικότητα ή το γεωγραφικό ιδιωματισμό του συγγραφέα. Για αυτό το λόγο χωρίσαμε αυτά τα χαρακτηριστικά σε δύο κατηγορίες:
  1. **Χαρακτηριστικά συχνότερων λέξεων ανά εθνικότητα**: Το συγκεκριμένο χαρακτηριστικό αποτυπώνει της περισσότερο χρησιμοποιημένες ανά συγγραφείς με την ίδια γεωγραφική προέλευση. Επέλεξα το παρόν χαρακτηριστικό καθώς οι διαφοροποιήσεις που υπάρχουν από χώρα σε χώρα στη χρήση της αγγλικής μπορεί να μας βοηθήσει στο να εξάγουμε γνώση από την εθνικότητα του συγγραφέα.
  2. **Χαρακτηριστικό συχνότητα χρήσης αργκό**: Σε αυτό το χαρακτηριστικό αναζήτησα τις ποιο γνωστές λέξεις αργκό που συναντάμε στις αγγλόφωνες χώρες. Δημιούργησα δηλαδή σύνολα με λέξεις αργκό ανά χώρα χωρίς όμως να υπάρχουν κοινές λέξεις μεταξύ των συνόλων. Η επιλογή του συγκεκριμένου χαρακτηριστικού βασίζεται στους γλωσσικούς ιδιωματισμούς που συναντώνται αλλά και της μοναδικότητας λέξεων αργό από χώρα σε χώρα.
- **Κλάση Γεωγραφικού ιδιωματισμού**: Με αυτό το χαρακτηριστικό προσδιορίζουμε την κλάση ως προς την οποίο θα γίνει η κατηγοριοποίηση κειμένου. Για λόγους απλότητας η τιμή της κλάσης παίρνει 5 τιμές από το 0 έως το 4 με την εξής λογική:
  - 0 όταν η επισημείωση του εγγράφου έχει την ετικέτα *‘US’*
  - 1 όταν η επισημείωση του εγγράφου έχει την ετικέτα *‘AUS’*
  - 2 όταν η επισημείωση του εγγράφου έχει την ετικέτα *‘CAN’*
  - 3 όταν η επισημείωση του εγγράφου έχει την ετικέτα *‘UK’*
  - 4 όταν η επισημείωση του εγγράφου έχει την ετικέτα *‘NNS’*

επέλεξα την χρήση ακεραίου αριθμού για τον προσδιορισμό της κλάσης για ταχύτερη

απόκριση του αλγορίθμου κατηγοριοποίησης.

id	**Χαρακτηριστικά**
	Λεξιλογικά Δεδομένα
	Χαρακτηριστικά βασισμένα στους χαρακτήρες
	Συνολικός αριθμός χαρακτήρων κειμένου (C)
1	Συνολικός αριθμός συμβόλων / C
2	Συνολικός αριθμός σημείων στίξης / C
3	Συνολικός αριθμός κενών χαρακτήρων / C
4	Συνολικός αριθμός κεφαλαίων χαρακτήρων / C
5	Συνολικός αριθμός αλφαβητικών χαρακτήρων / C
6	Συνολικός αριθμός ψηφίων / C
15-40	Συχνότητα εμφάνισης γραμμάτων (A-Z, 26 χαρακτηριστικά)
41-54	Συχνότητα εμφάνισης ειδικών χαρακτήρων (~, @, #, $, ^, &, *, -, _, =, +,>, <, [,], {, }, /, \, |, 14 χαρακτηριστικά)
	Χαρακτηριστικά βασισμένα στις λέξεις
	Συνολικός αριθμός λέξεων (M)
7	Συνολικός αριθμός μικρών λέξεων (<4 χαρακτήρες) / M
8	Συνολικός αριθμός χαρακτήρων στις λέξεις / C
9	Μέσο μήκος λέξης
13	Άπαξ λεγόμενα (Συνχότητα των λέξεων που εμφανίζονται μια φορά)
14	Άπαξ δυσλεγόμενα (Συχνότητα των λέξεων που εμφανίζονται δύο φορές)
10	Μέσο μήκος πρότασης / M
11	Μέσο μήκος πρότασης / C
12	Διαφορετικές Λέξεις / M
	Συντακτικά Χαρακτηριστικά
55-67	Συχνότητα σημείων στίξης (“,”, “.”, “?”, “!”, “:”, “;”, “ ' ”, “ ” ”)
	Χαρακτηριστικά με βάση το περιεχόμενο
	Χαρακτηριστικά συχνότερων λέξεων ανά εθνικότητα
68	Συχνότητα χρήσης των most American used words
70	Συχνότητα χρήσης των most Austrelian used words
69	Συχνότητα χρήσης των most British used words
71	Συχνότητα χρήσης των most Canadian used words
72	Συχνότητα χρήσης των most NNS used words
	Χαρακτηριστικά συχνότητα χρήσης αργκό
73	Συχνότητα χρήσης αμερικάνικης αργκό
74	Συχνότητα χρήσης αγγλικής αργκό
76	Συχνότητα χρήσης καναδικής αργκό
75	Συχνότητα χρήσης αυστραλιανής αργκό
77	Κλάση γεωγραφικού ιδιωματισμού (‘0’, ’1’, ’2’, ’3’, ’4’ για ‘US’, ‘AUS’, ‘CAN, ‘UK’, ‘NNS’ αντίστοιχα)

Πίνακας 5: Σύνολο χαρακτηριστικών

3 Εξαγωγή Χαρακτηριστικών

Η εξαγωγή των χαρακτηριστικών έγινε με υλοποίηση σε Python[5] (Python 2.7.9) και με τη βοήθεια των βιβλιοθηκών που προσφέρει το NLTK 3.1. Αρχικά υλοποιήσαμε το script με όνομα rem_empty.py το οποίο σαρώνει το αρχείο με το σύνολο των κειμένων και εάν εντοπίσει πεδίο κειμένου που είναι κενό τότε διαγράφει την γραμμή (instance). Υλοποιήσαμε ένα script το οποίο στην ουσία "σαρώνει" το αρχείο με το προ επεξεργασμένο και επισημειωμένο σύνολο των κειμένων και υπολογίζει για το κάθε κείμενο τις τιμές των 75 χαρακτηριστικών που περιγράψαμε παραπάνω. Το αρχείο .csv με το σύνολο το επισημειωμένων κειμένων έχει κωδικοποίηση unicode την οποία δεν αναγνωρίζει η βασική βιβλιοθήκη[6] της Python για επεξεργασία αρχείων csv. Για αυτό το λόγο για την επεξεργασία του αρχείου csv εισάγουμε βιβλιοθήκη[7] με δυνατότητα υποστήριξης κωδικοποίησης των αρχείων csv. Από τις βιβλιοθήκες του NLTK χρησιμοποιούμε τις εξής:

`nltk`

`tokenize`[8] Για τον χωρισμό του αρχικού string/εγγράφου σε μικρότερα string για υπολογισμό και εξαγωγή χαρακτηριστικών

`regexp`[9] Χωρίζει το αρχικό string σε υπό string με βάσει κανονικές εκφράσεις που λαμβάνει ως όρισμα. π.χ με το συγκεκριμένο module μετράμε τη συχνότητα εμφάνησης γραμμάτων/συμβόλων κλπ.

`punkt`[10] Με αυτό το module γίνεται χωρισμός του κειμένου σε προτάσεις το οποίο είναι απαραίτητο για τον υπολογισμό χαρακτηριστικών. Στην συγκεκριμένη περίπτωση χρησιμοποιούμε το αρχείο για αγγλικό κείμενο **`tokenizers/punkt/english.pickle`**

Με τη χρήση της αντίστοιχης βιβλιοθήκης[11] της Python ο κώδικας της υλοποίησης εξάγει ένα αρχείο .arff με τα χαρακτηριστικά για για κάθε κείμενο του συνόλου που εισάγαμε. Με το συγκεκριμένο αρχείο θα πραγματοποιήσουμε τη διαδικασία της κατηγοριοποίησης στο περιβάλλον Weka.

4 Επιλογή Χαρακτηριστικών (Feature Selection)

Η διαδικασία της επιλογής χαρακτηριστικών προηγείται της κατηγοριοποίησης και αποτελεί τη διαδικασία αξιολόγησης των χαρακτηριστικών. Για αυτή τη διαδικασία χρησιμοποιήθηκε ο αλγόριθμος ReliefF. Στο παρακάτω πίνακα παραθέτεται τα αποτελέσματα.

A/A	ReliefF score	Όνομα Χαρακτηριστικού

5 https://www.python.org/downloads/release/python-279/
6 csv 1.0 https://pypi.python.org/pypi/csv/1.0
7 unicodecsv 0.13.0 https://pypi.python.org/pypi/unicodecsv/0.13.0
8 http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize
9 http://www.nltk.org/.../.tokenize.regexp
10 http://www.nltk.org/.../.tokenize.punkt
11 arff 0.9 https://pypi.python.org/pypi/arff/0.9
liac-arff 2.1.0 https://pypi.python.org/pypi/liac-arff/2.1.0

1	0.00329917	Συνολικός αριθμός κεφαλαίων χαρακτήρων / C
2	0.00263219	Συνολικός αριθμός μικρών λέξεων (<4 χαρακτήρες) / M
3	0.00190155	Συνολικός αριθμός κενών χαρακτήρων / C
4	0.00181929	Μέσο μήκος λέξης
5	0.00175353	Συνολικός αριθμός αλφαβητικών χαρακτήρων / C
6	0.00149185	Συνολικός αριθμός σημείων στίξης / C
7	0.00132278	Συνολικός αριθμός χαρακτήρων στις λέξεις / C
8	0.00108097	Συνολικός αριθμός συμβόλων / C
9	0.00095305	avg_sentences/char
10	0.00086441	Συνολικός αριθμός ψηφίων / C
11	0.00072979	Άπαξ λεγόμενα
12	0.00065478	Άπαξ δυσλεγόμενα
13	0.00064417	Μέσο μήκος πρότασης / M
14	0.00049383	Συχνότητα χρήσης των most British used words
15	0.00046527	Διαφορετικές Λέξεις / M
16	0.00044018	Συχνότητα χρήσης των most NNS used words
17	0.00038819	Συχνότητα χρήσης των most American used words
18	0.00026978	Συχνότητα χρήσης των most Canadian used words
19	0.00015981	Συχνότητα εμφάνισης σημείου στίξης quotationmark1
20	0.00015981	Συχνότητα εμφάνισης σημείου στίξης exclamationmark
21	0.00015981	Συχνότητα εμφάνισης σημείου στίξης quotationmark2
22	0.00015981	Συχνότητα εμφάνισης σημείου στίξης semicolon
23	0.00015981	Συχνότητα εμφάνισης σημείου στίξης comma
24	0.00015981	Συχνότητα εμφάνισης σημείου στίξης fullstop
25	0.00015981	Συχνότητα εμφάνισης σημείου στίξης questionmark
26	0.00015981	Συχνότητα εμφάνισης σημείου στίξης colon
27	0.00010395	Συχνότητα χρήσης των most Austraelian used words
28	0.00009802	Συχνότητα σημείου στίξης bracket4
29	0.00009802	Συχνότητα εμφάνισης σημείου στίξης verticalbar
30	0.00009585	Συχνότητα εμφάνισης συμβόλου hash
31	0.00009585	Συχνότητα εμφάνισης συμβόλου at
32	0.00009585	Συχνότητα εμφάνισης συμβόλου tilde
33	0.00009585	Συχνότητα εμφάνισης συμβόλου percent
34	0.00009585	Συχνότητα εμφάνισης συμβόλου dollar

35	0.00009585	Συχνότητα εμφάνισης συμβόλου dash
36	0.00009585	Συχνότητα εμφάνισης συμβόλου asterisk
37	0.00009585	Συχνότητα εμφάνισης συμβόλου caret
38	0.00009585	Συχνότητα εμφάνισης συμβόλου ampersand
39	0.00008196	Συχνότητα εμφάνισης γράμματος “h”
40	0.00008196	Συχνότητα εμφάνισης γράμματος “j”
41	0.00008196	Συχνότητα εμφάνισης γράμματος “i”
42	0.00008196	Συχνότητα εμφάνισης γράμματος “f”
43	0.00008196	Συχνότητα εμφάνισης γράμματος “g”
44	0.00008196	Συχνότητα εμφάνισης γράμματος “c”
45	0.00008196	Συχνότητα εμφάνισης γράμματος “e”
46	0.00008196	Συχνότητα εμφάνισης γράμματος “d”
47	0.00008196	Συχνότητα εμφάνισης γράμματος “l”
48	0.00008196	Συχνότητα εμφάνισης γράμματος “a”
49	0.00008196	Συχνότητα εμφάνισης γράμματος “b”
50	0.00008196	Συχνότητα εμφάνισης γράμματος “k”
51	0.00008196	Συχνότητα εμφάνισης γράμματος “x”
52	0.00008196	Συχνότητα εμφάνισης γράμματος “m”
53	0.00008196	Συχνότητα εμφάνισης γράμματος “w”
54	0.00008196	Συχνότητα εμφάνισης γράμματος “u”
55	0.00008196	Συχνότητα εμφάνισης γράμματος “n”
56	0.00008196	Συχνότητα εμφάνισης γράμματος “t”
57	0.00008196	Συχνότητα εμφάνισης γράμματος “z”
58	0.00008196	Συχνότητα εμφάνισης γράμματος “v”
59	0.00008196	Συχνότητα εμφάνισης γράμματος “y”
60	0.00008196	Συχνότητα εμφάνισης γράμματος “s”
61	0.00008196	Συχνότητα εμφάνισης γράμματος “o”
62	0.00008196	Συχνότητα εμφάνισης γράμματος “p”
63	0.00008196	Συχνότητα εμφάνισης γράμματος “r”
64	0.00008196	Συχνότητα εμφάνισης γράμματος “q”
65	0.00003673	Συχνότητα εμφάνισης σημείου στίξης bracket1
66	0.00003673	Συχνότητα εμφάνισης σημείου στίξης bracket3
67	0.00003673	Συχνότητα εμφάνισης συμβόλου less
68	0.00003673	Συχνότητα εμφάνισης σημείου στίξης bracket2

69	0.00003673	Συχνότητα εμφάνισης συμβόλου plus
70	0.00003673	Συχνότητα εμφάνισης συμβόλου greater
71	0.00003673	Συχνότητα εμφάνισης συμβόλου dash1
72	0.00003673	Συχνότητα εμφάνισης συμβόλου equals_sign
73	0.00002534	Συχνότητα χρήσης αμερικάνικης αργκό
74	0.00000757	Συχνότητα χρήσης αυστραλιανής αργκό
75	0.00000591	Συχνότητα χρήσης καναδικής αργκό
76	0.00000477	Συχνότητα χρήσης αγγλικής αργκό

Πίνακας 6: Feature Ranking

5 Πειραματικά αποτελέσματα κατηγοριοποίησης

Μετά την αξιολόγηση των χαρακτηριστικών γίνεται η δημιουργία υπο-ομάδων χαρακτηριστικών. Για να την επίτευξη των καλύτερων δυνατών αποτελεσμάτων εκτελούμε τους αλγόριθμους κατηγοριοποίησης σε ένα σύνολο υποσυνόλων των χαρακτηριστικών, δηλαδή για 10 χαρακτηριστικά, 20, 30 κ.ο.κ. Προφανώς η επιλογή των χαρακτηριστικών κάθε φορά για την κάθε υπο-ομάδα γίνεται βάσει της κατάταξης που δημιουργήθηκε και περιγράφεται στην προηγούμενη ενότητα.

Κατηγοριοποιητής	top10	top20	top30	top40	top50	top60	all
J48	52.28	**53.88**	52.24	51.23	50.31	51.26	51.30
MLP	47.17	51.16	**51.20**	51.11	50.07	49.36	45.01
RandomTree	**55.58**	55.20	50.66	46.01	41.53	40.62	40.52
REPTree	50.45	53.18	**53.22**	52.87	52.01	52.87	52.88
RBFNetwork	46.70	**49.30**	48.89	48.00	48.53	47.91	48.44
Bagging	57.96	**61.87**	59.61	58.34	57.50	58.34	58.35
Boosting	46.71	47.62	47.62	47.62	**47.78**	47.62	47.62
IBk	**56.34**	49.08	46.52	41.57	39.08	39.05	38.88

Πίνακας 7: Αποτελέσματα κατηγοριοποίησης

Με βάση τον πίνακα των αποτελεσμάτων καλύτερος αλγόριθμος είναι ο Bagging καθώς στο σύνολο των 20 χαρακτηριστικών πετυχαίνει 61.87 ποσοστό επιτυχίας. Ο ίδιος αλγόριθμος σημειώνει καλύτερα αποτελέσματα σε όλα τα σύνολα των χαρακτηριστικών. Παρατηρείτε επίσης ότι καλύτερα αποτελέσματα σημειώνονται στα σύνολα με μικρό αριθμό χαρακτηριστικών εκτός του αλγορίθμου Boosting που σημειώνει μεγαλύτερο ποσοστό στο σύνολο των 50 χαρακτηριστικών.

Ενδιαφέρον παρουσιάζει ότι το ποσοστό του κάθε αλγορίθμου παρουσιάζει μικρές διαφορές από

σύνολο σε σύνολο χαρακτηριστικών. Οι αυξομειώσεις στις περισσότερες περιπτώσεις είναι της τάξης του 2%. Εξαίρεση αποτελούν οι αλγόριθμοι IBk και Random Tree όπου εκεί οι διαφορές μεγαλύτερου-μικρότερου ποσοστού είναι 17.46% και 15.06% αντίστοιχα. Ο αλγόριθμος με το μικρότερο ποσοστό είναι ο IBk στο σύνολο όλων των χαρακτηριστικών με τιμή 38.88%

Κεφάλαιο 5 – Συμπεράσματα

Στόχος της παρούσας διπλωματικής εργασίας ήταν να υλοποιηθεί ένα σύστημα το οποίο να δίνει λύση στο πρόβλημα της κατηγοριοποίησης κειμένων με βάση τον γεωγραφικό ιδιωματισμό του συγγραφέα. Το συγκεκριμένο πρόβλημα είναι απ τη φύση του πιο δύσκολο καθώς ξεπερνάει το επίπεδο δυσκολίας της απλής κατηγοριοποίησης κειμένου. Αυτό συμβαίνει γιατί για τη συγκεκριμένη κατηγοριοποίηση πρέπει να ληφθεί υπόψιν ένας μεγάλος αριθμός χαρακτηριστικών για τα οποία πολλές φορές παρουσιάζεται δυσκολία στον υπολογισμό και την εξαγωγή τους. Επιπλέον αξίζει να ληφθεί υπόψιν ότι για την συγκεκριμένη εργασία υπήρχε μικρός αριθμός βιβλιογραφικών αναφορών. Η έως τώρα μελέτες στο πεδίο της αυτόματης αναγνώρισης συγγραφέα με χρήση κατηγοριοποίησης κειμένου εστίαζαν στην κατηγοριοποίηση ως προς το φύλο και την ηλικία του συγγραφέα. Έτσι οι μελέτες στην κατηγοριοποίηση κειμένου με βάση το γεωγραφικό στοιχείο ήταν περιορισμένες κάνοντας την διαδικασία της επιλογής και εξαγωγής χαρακτηριστικών αρκετά δύσκολη.

Από την βιβλιογραφική μελέτη παρατηρούμε ότι σε αντίστοιχες εργασίες με κατηγοριοποίηση ως προς το φύλλο ή την ηλικία του συγγραφέα τα αποτελέσματα της εργασίας μπορούν να λάβουν υψηλά ποσοστά επιτυχίας. Στην παρούσα διπλωματική τα αποτελέσματα της κατηγοριοποίησης είναι χαμηλότερα όμως αν λάβουμε υπόψιν τον μικρό αριθμό εργασιών στην εξαγωγή χαρακτηριστικών βάσει του γεωγραφικού ιδιωματισμού τότε τα αποτελέσματα της παρούσας εργασία μπορούν να χαρακτηριστούν ως ικανοποιητικά.

Για την εξαγωγή των χαρακτηριστικών δουλέψαμε με τον εξής τρόπο. Επιλέξαμε μια σειρά χαρακτηριστικών που χρησιμοποιούνται ευρέως στην κατηγοριοποίηση κειμένου. Στη συνέχεια επιλέξαμε χαρακτηριστικά κατάλληλα για την κατηγοριοποίηση κειμένου βάσει του γεωγραφικού ιδιωματισμού, η εξαγωγή αυτών των χαρακτηριστικών μπορούν να αποτελέσουν αφορμή για ευρύτερη έρευνα πάνω στο πεδίο και βελτιστοποίηση της εξαγωγής χαρακτηριστικών για κατηγοριοποίηση με βάση γεωγραφικούς ιδιωματισμούς.

Για τα πειράματα κατηγοριοποίησης επιλέξαμε να χρησιμοποιήσουμε πλήθος αλγορίθμων έτσι ώστε να αναδείξουμε αυτόν με τα καλύτερα αποτελέσματα. Επίσης για την βελτιστοποίηση των αποτελεσμάτων προχωρήσαμε και στην επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου ReliefF. Τέλος τα πειράματα κατηγοριοποίησης τα εφαρμόσαμε σε πλήθος υποσυνόλων τωω χαρακτηριστικών.

Επίλογος

Στην παρούσα εργασία προσπαθήσαμε να προσεγγίσουμε τη λύση του προβλήματος της αυτόματης αναγνώρισης συγγραφέα με βάση τον γεωγραφικό ιδιωματισμό. Σε σχέση με την πλειοψηφία των εργασιών του συγκεκριμένου ερευνητικού πεδίου που καταπιάνονται με την αυτόματη αναγνώριση συγγραφέα ως προς το φύλλο ή την ηλικία, η παρούσα διπλωματική διαφοροποιείται. Ακόμα η παρούσα εργασία θα μπορούσε, στον βαθμό που της αναλογεί να συμβάλει στην περαιτέρω έρευνα στο πεδίο της αυτόματης αναγνώρισης συγγραφέα με βάση γεωγραφικούς ιδιωματισμούς. Όσο ο όγκος των δεδομένων στον παγκόσμιο ιστό μεγαλώνει τόσο η εξαγωγή της πληροφορίας γίνεται πιο δύσκολη. Έτσι στοιχεία όπως ο γεωγραφικός ιδιωματισμός του συγγραφέα μπορούν να δώσουν άλλες διαστάσεις στο πρόβλημα της κατηγοριοποίησης κειμένου καθώς και δημιουργήσει νέες δυνατότητες στην αυτόματη αναγνώριση συγγραφέα.

Βιβλιογραφία

Aravantinou, C., Simaki, V., Mporas, I., & Megalooikonomou, V. (2015). Gender Classification of Web Authors Using Feature Selection and Language Models. In Speech and Computer (pp. 226-233). Springer International Publishing.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Caropreso, M. F., Matwin, S., & Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text databases and document management: Theory and practice*, 78-102.

Classification of Web Authors Using Feature Selection and Language Models. In Speech and Computer (pp. 226-233). Springer International Publishing.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, *2*, 265-292.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, *29*(2-3), 103-130.

Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval: data structures and algorithms.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, *28*(2), 87-106.

Keyling, Till; Jünger, Jakob (2013). Facepager (Version, f.e. 3.3). An application for generic data retrieval through APIs. Source code available from https://github.com/strohne/Facepager.

Koppel, M., & Schler, J. (2004, July). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning* (p. 62). ACM.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401-412.

Koppel, M., Schler, J., & Zigdon, K. (2005, August). Determining an author's native language by mining a text for errors. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 624-628). ACM.

Margaret H. Dunham: Data Mining Εισαγωγή και Προηγμένα Θέματα Εξωρυξης Γνώσης από Δεδομένα.

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M. (2014, January). Age and gender identification in social media. In *Proceedings of CLEF 2014 Evaluation Labs* (pp. 1129-1136).

Natural Language Toolkit. (2016, March 3). In *Wikipedia, The Free Encyclopedia*. Retrieved 01:05, March 20, 2016, from https://en.wikipedia.org/w/index.php?title=Natural_Language_Toolkit&oldid=708009539

Python. (2015, Δεκεμβρίου 22). *Βικιπαίδεια, Η Ελεύθερη Εγκυκλοπαίδεια*. Ανακτήθηκε 01:06, Μαρτίου 20, 2016 από το //el.wikipedia.org/w/index.php?title=Python&oldid=5600345.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sebastiani, F. (2005). Text Categorization.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

Stein, B., & zu Eissen, S. M. (2007). Intrinsic Plagiarism Analysis with Meta Learning. *PAN*, *276*.

Weka (machine learning). (2016, March 11). In *Wikipedia, The Free Encyclopedia*. Retrieved 01:07, March 20, 2016, from https://en.wikipedia.org/w/index.php?title=Weka_(machine_learning)&oldid=709492082

Weka (Μηχανική Μάθηση). (2016, Μαΐου 29). *Βικιπαίδεια, Η Ελεύθερη Εγκυκλοπαίδεια*. Ανακτήθηκε 16:36, Ιουλίου 19, 2016 από το //el.wikipedia.org/w/index.php?title=Weka_(%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%9C%CE%AC%CE%B8%CE%B7%CF%83%CE%B7)&oldid=5873222.

Yang, Y. (2001, September). A study of thresholding strategies for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 137-145). ACM.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3), 378-393.

Zobel, J., & Moffat, A. (1998, April). Exploring the similarity space. In *ACM SIGIR Forum* (Vol. 32, No. 1, pp. 18-34). ACM.

Zu Eissen, S. M., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. In *Advances in data analysis* (pp. 359-366). Springer Berlin Heidelberg.

Αθανασοπούλου, Ευαγγελία-Ελένη. "Εφαρμογές της μηχανικής μάθησης στην κατηγοριοποίηση κεμένου." (2006).

Αραβαντινού, Χ. (2015). Ανάπτυξη μεθόδων αυτόματης κατηγοριοποίησης κειμένων προσανατολισμένων στο φύλο (Doctoral dissertation).

Εξόρυξη δεδομένων. (2016, Φεβρουαρίου 6). *Βικιπαίδεια, Η Ελεύθερη Εγκυκλοπαίδεια*. Ανακτήθηκε 23:48, Μαρτίου 19, 2016 από το //el.wikipedia.org/w/index.php?title=%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD&oldid=5671036.

Χαραλαμπόπουλος, Ι. Δ. (2011). Κατηγοριοποίηση στον Παγκόσμιο Ιστό και στο περιβάλλον WEKA.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, *53*(1-2), 23-69.