Ιονιο Πανεπιστημιο

Τμημα Πληροφορικης

– Πτυχιακή Εργασία –

Αυτόματη Επιλογή και Ταξινόμηση Δεδομένων Εισόδου Βάσει Χρησιμότητας για Αναλυτικές Εργασίες Μεγάλου Όγκου Δεδομένων

Κοτσαγιαννίδης Πέτρος
Επιβλέπων: Θεμιστοκλής Έξαρχος

16 Ιανουαρίου 2022

Επιβλέπων

Θεμιστοκλής Έξαρχος, *Επίκουρος καθηγητής,*
Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Θεμιστοκλής Έξαρχος, *Επίκουρος καθηγητής,*
Ιόνιο Πανεπιστήμιο
Μιχαήλ Στεφανιδάκης, *Αναπληρωτής καθηγητής,*
Ιόνιο Πανεπιστήμιο
Θεόδωρος Ανδρόνικος, *Αναπληρωτής καθηγητής,*
Ιόνιο Πανεπιστήμιο

Περίληψη

Καθ΄ ολη την διάρκεια της εξέλιξης του ανθρώπινου πολιτισμού, ένα από τα πιο βασικά στοιχεία που συντελούν στην ανάπτυξη του είναι τα δεδομένα, και κατ΄ επέκταση η πληροφορία. Πάντοτε η πρόοδος βασίζεται στην μελέτη, επεξεργασία και αξιοποίηση πληροφορίας που έχει αποθηκευτεί, καθώς χωρίς την δυνατότητα αποθήκευσης της δεν θα ήταν εφικτή η διατήρηση της στο πέρασμα των χρόνων. Το ίδιο ακριβώς ισχύει και στην εποχή του διαδικτύου, απλός πλέον ο όγκος της πληροφορίας και των δεδομένων είναι τεράστιος και συνεπώς υπάρχουν πολλά εμπόδια κατά την επεξεργασία και αξιοποίηση τους. Μία πρόκληση πλέον είναι να δημιουργηθούν εργαλεία τα οποία μπορούν αποδοτικά και σε σχετικά μικρό χρονικό διάστημα να εξάγουν όσο το δυνατόν περισσότερη γνώση από όσο το δυνατόν μεγαλύτερο όγκο δεδομένων. Έχουν αναπτυχθεί πολλές τεχνολογίες για να επιτευχθεί η γρήγορη επεξεργασία μεγάλου όγκου δεδομένων, αλλά η κύρια πρόκληση είναι η διαχρονικότητά τους, καθώς η ποσότητα των διαθέσιμων δεδομένων αυξάνεται με εκθετικούς ρυθμούς και τα εργαλεία καθίστανται απαρχαιωμένα σε μικρό χρονικό διάστημα.

Στην παρούσα εργασία προτείνεται μία μέθοδος που επιχειρεί να μειώσει τον απαιτούμενο χρόνο εφαρμογής ενός τελεστή σε ένα μεγάλο σύνολο δεδομένων κειμένου. Σκοπός είναι να επιτευχθεί η πρόβλεψη του αποτελέσματος ενός αλγορίθμου, που εφαρμόζεται σε δεδομένα κειμένου, για ένα μεγάλο σύνολο χωρίς να χρειάζεται να εισάγουμε σε αυτόν όλα τα δεδομένα. Για να είναι εφικτή η πρόβλεψη απαιτούνται δύο στοιχεία, ο βαθμός ομοιότητας ('Similarity') όλων των εγγραφών του συνόλου μεταξύ τους και το αποτέλεσμα του τελεστή για ένα μικρό ποσοστό του συνόλου. Με βάση αυτά τα δύο προβλέπεται το αποτέλεσμα του για τις υπόλοιπες εγγραφές, και στόχος είναι να επιτυγχάνεται με μεγάλη ακρίβεια. Θεωρητικά δεν υπάρχει κάποιος περιορισμός στο είδος του αλγορίθμου που θα εφαρμοστεί η μέθοδος, παρά μόνο να δέχεται σαν είσοδο ένα κείμενο σε φυσική γλώσσα. Για την δοκιμή της μεθόδου επιλέχθηκε ένας αλγόριθμος αναγνώρισης γλώσσας, όπου επιστρέφει την γλώσσα που είναι γραμμένο το κείμενο στο οποίο έχει εφαρμοστεί, και δοκιμάστηκε για διαφόρων ειδών κείμενα σε πέντε λατινογενείς γλώσσες. Για τον βαθμό ομοιότητας δοκιμάστηκαν τρία διαφορετικά μέτρα υπολογισμού, η Ευκλείδεια απόσταση, η ομοιότητα "Cosine" και το "Word Movers Distance".

Μέσω των δοκιμών φάνηκε πως η μέθοδος της πρόβλεψης της εξόδου ενός αλγορίθμου για ένα σύνολο δεδομένων κειμένου με βάση την ομοιότητά τους είναι ως ένα βαθμό εφικτή, καθώς με την χρήση της ομοιότητας Cosine σαν μέτρο απόστασης επιτεύχθηκε, στην καλύτερη περίπτωση, ακρίβεια της τάξεως του 93%. Σε μία μέση περίπτωση, όπου πρέπει να προβλεφθεί η έξοδος για το 97% του συνόλου, η μέση ακρίβεια είναι της τάξεως του 87%. Όσον αφορά την χρονική πολυπλοκότητα, η μέθοδος δεν κατάφερε να την μειώσει για τον συγκεκριμένο αλγόριθμο αλλά ενδεχομένως να είναι εφικτό για κάποιον πιο πολύπλοκο τελεστή.

Συμπερασματικά, η επεξεργασία της φυσικής γλώσσας και η εξαγωγή γνώσης από δεδομένα κειμένου είναι δύο πολύ περίπλοκες διαδικασίες αλλά σίγουρα πολύ αναγκαίες, καθώς στις μέρες μας που το διαδίκτυο είναι προσβάσιμο από όλους και ο καθένας μπορεί να εκφράζει την γνώμη του μέσα από αυτό, υπάρχει ατελείωτη πληροφορία που θα μπορούσε να φανεί χρήσιμη σε πολλές διαφορετικές περιπτώσεις. Στην παρούσα εργασία προτάθηκε μια μεθοδολογία η οποία θα μπορούσε υπό συνθήκες να μειώσει τον χρόνο που απαιτείται από έναν αλγόριθμο

για να εφαρμοστεί σε ένα μεγάλο σύνολο δεδομένων, και εκ του αποτελέσματος φάνηκε ότι με περιορισμούς ήταν εν μέρει επιτυχής. Σίγουρα υπάρχουν πολλά περιθώρια βελτίωσης, αλλά έγινε το πρώτο βήμα για την δοκιμή της αποτελεσματικότητάς της.

Περιεχόμενα

Κατάλογος Πινάκων

Κεφάλαιο Α΄

Εισαγωγή

Α΄.1 Διαχείριση μεγάλου όγκου δεδομένων και προκλήσεις

Τα τελευταία χρόνια, η ανθρωπότητα έχει αντιληφθεί την μεγάλη αξία των δεδομένων και της γνώσης που μπορείς να εξάγεις από αυτά, γι΄ αυτό και συνεχής έρευνες προσπαθούν να ανακαλύψουν τρόπους ώστε αυτά να αξιοποιηθούν. Πλέον, έχουν απαντηθεί πολλά ερωτήματα που αφορούν την επεξεργασία, την ανάλυση και την εκμετάλλευση δεδομένων και η νέα πρόκληση είναι να βρεθούν αποδοτικές λύσεις για την εφαρμογή αντίστοιχων τεχνικών σε μεγάλης κλίμακας σύνολα δεδομένων.

Μιας και δεν γίνεται να τεθεί κάποιο όριο για το ποια σύνολα δεδομένων μπορούν να χαρακτηριστούν μεγάλα και ποια όχι, θα μπορούσαμε να ορίσουμε μεγάλο τον όγκο των δεδομένων όταν αυτά δεν μπορούν να επεξεργαστούν με τα κλασσικά εργαλεία ανάλυσης ή όταν το κόστος του χρόνου που χρειάζεται για να αναλυθούν με αυτά, υπερβαίνει το κέρδος της ανάλυσης τους.

Με την ραγδαία ανάπτυξη του διαδικτύου, ο όγκος των περίπλοκων δεδομένων που αποθηκεύονται και χρήζουν επεξεργασίας συνεχώς πολλαπλασιάζεται, με αποτέλεσμα να υπάρχει αδιάκοπη ανάγκη για εύρεση νέων, πιο αποδοτικών μεθόδων ανάλυσης τους. Κάποια από τα κύρια προβλήματα που καλούνται να αντιμετωπίσουν οι ερευνητές στον τομέα της διαχείρισης μεγάλου όγκου δεδομένων σύμφωνα με το [1] είναι i) η ακανόνιστη μορφή των δεδομένων και οι τυχαίες συσχετίσεις μεταξύ τους, ii) το μεγάλο υπολογιστικό κόστος ανάλυσης τους και iii) ο συνδυασμός πολλών δεδομένων, από πολλές διαφορετικές πηγές σε διαφορετικά χρονικά διαστήματα με την χρήση πολλών διαφορετικών εργαλείων.

Η ανάλυση μεγάλου όγκου δεδομένων είναι μία διαδικασία η οποία βρίσκει εφαρμογή σε πάρα πολλούς διαφορετικούς κλάδους, με σκοπό την βελτιστοποίηση των παροχών και την λήψη αποφάσεων για μελλοντικές ενέργειες. Στον τραπεζικό τομέα αξιοποιούνται μεγάλα σύνολα δεδομένων για την αποφυγή εξαπάτησης των πελατών, την βελτίωση των προσφορών ανάλογα με τις ανάγκες τους, την κατηγοριοποίηση τους με βάση κάποια χαρακτηριστικά, την ανάλυση ρίσκου μελλοντικών επενδύσεων κ.ά. [2] . Στον τομέα της υγείας, η ανάλυση δεδομένων χρησιμοποιείται για την πρόωρη διάγνωση ασθενειών, την εύρεση παραγόντων που

συμβάλλουν στην εξέλιξη κάποιας ασθένειας , την βελτίωση των νοσοκομειακών παροχών, την εξατομικευμένη θεραπεία κ.ά. [3], [4]. Στον κλάδο της αστρονομίας, η ανάλυση μεγάλου όγκου δεδομένων χρησιμοποιείται σε συστήματα που προβλέπουν την μορφολογία κάποιου γαλαξία, στην κατηγοριοποίηση διαστημικών αντικειμένων, στην αναγνώριση υπερκαινοφανών αστέρων, την πρόβλεψη αποστάσεων κ.ά. [5].

Είναι ξεκάθαρο πως η εξαγωγή γνώσης από δεδομένα δεν είναι μια διαδικασία που περιορίζεται σε συγκεκριμένους τομείς, αλλά αντίθετα μπορεί να χρησιμοποιηθεί σε πάρα πολλές περιπτώσεις και να αποτελέσει βασικό κομμάτι της εξέλιξης ενός οργανισμού, αν συλλεχθούν τα κατάλληλα δεδομένα. Τα παραπάνω παραδείγματα, δείχνουν πως 3 τελείως διαφορετικοί κλάδοι μπορούν να λύσουν βασικά τους προβλήματα μέσα από την αξιοποίηση των σωστών δεδομένων, και να αποκομίσουν πολύ σημαντικές πληροφορίες που τους βοηθούν στην καλύτερη λήψη αποφάσεων. Τα δεδομένα μπορούν να φανούν χρήσιμα σε όλους τους τομείς, αρκεί να δοθεί βάση στην ποιότητα τους.

Α΄.2 Επιλογή κατάλληλων δεδομένων

Είναι γεγονός, πως το μεγαλύτερο μέρος των εργαλείων που αφορούν την διαχείριση μεγάλου όγκου δεδομένων βασίζεται κυρίως στο μέγεθος του συνόλου των δεδομένων και όχι στα δεδομένα καθ΄ εαυτά. Κυριαρχεί στις επιχειρήσεις η άποψη πως όσα περισσότερα δεδομένα διαθέτεις και αναλύεις τόσο μεγαλύτερο το όφελος. Τελευταία όμως έχουν προταθεί και απόψεις που δεν παρουσιάζουν την ποσότητα των δεδομένων σαν πιο κρίσιμο παράγοντα επίλυσης προβλημάτων, αλλά την ποιότητα. Ίσως είναι πιο ωφέλιμο να επεξεργάζονται και να αναλύονται μικρότερα αλλά φιλτραρισμένα, με βάση την χρησιμότητα, σύνολα δεδομένων από το να αναλύονται τεράστια, ακανόνιστα και πολλές φορές αχρείαστα σύνολα. Πως όμως μπορούν να επιλεχθούν τα κατάλληλα δεδομένα για κάποια διεργασία;

Τα προβλήματα που πηγάζουν από τα δεδομένα και πρέπει να αντιμετωπιστούν για να επιλεχθούν τα κατάλληλα είναι πολυδιάστατα και συναντώνται κυρίως στις πηγές των δεδομένων οι οποίες είναι πάρα πολλές, με διαφορετική δομή και για διαφορετικούς σκοπούς, καθώς και στα ίδια τα δεδομένα τα οποία είναι αυθαίρετα, πολύπλοκα, χρονοβόρα στην επεξεργασία τους και ανανεώνονται ταχύτατα. Θα ήταν όμως πολύ ωφέλιμο να μπορούσαμε με βάση κάποια κριτήρια να επιλέγαμε το κατάλληλο υποσύνολο δεδομένων για μια διεργασία, μέσα από έναν τεράστιο αριθμό συνόλων δεδομένων από πολλές διαφορετικές πηγές. Έτσι δεν θα χρειαζόταν να υλοποιήσουμε την διαδικασία για το σύνολο των διαθέσιμων δεδομένων, αλλά μόνο για ένα μικρό υποσύνολο που πιθανώς να εμπεριέχει όλη την πληροφορία που χρειαζόμαστε.

Το να μπορούμε να προβλέπουμε το αποτέλεσμα ενός αλγορίθμου για ένα τεράστιο σύνολο δεδομένων ή για πολλά τέτοια σύνολα είναι μια διαδικασία που θα μπορούσε να φανεί πολύ χρήσιμη σε αναλυτικές εργασίες μεγάλου όγκου δεδομένων. Υποθέτοντας πως υπάρχουν χαρακτηριστικά σε ένα σύνολο δεδομένων τα οποία επηρεάζουν σημαντικά το αποτέλεσμα ενός αλγορίθμου που εφαρμόζεται σε αυτό. Ποια είναι αυτά τα χαρακτηριστικά που έχουν μεγάλο αντίκτυπο στο αποτέλεσμα κάποιου αλγορίθμου;Πως θα μπορούσαμε να τα αξιοποιήσουμε για να μαντέψουμε την έξοδο του αλγορίθμου; Είναι λοιπόν πρόκληση το να αναπτύξουμε τεχνο-

λογίες οι οποίες μπορούν να προβλέπουν με μεγάλη ακρίβεια την έξοδο ενός αλγορίθμου για ένα σύνολο δεδομένων, με βάση χαρακτηριστικά των δεδομένων που επηρεάζουν σημαντικά το αποτέλεσμα του.

Α΄.3 Δεδομένα φυσικής γλώσσας και ανάγκη για αυτόματη επιλογή

Με την ευρεία διάδοση το μέσων κοινωνικής δικτύωσης και άλλων εφαρμογών, έχουν αυξηθεί υπερβολικά και τα σύνολα δεδομένων σε φυσική γλώσσα τα οποία πρέπει να αναλυθούν και να εξαχθεί γνώση από αυτά. Η εξόρυξη γνώσης από δεδομένα κειμένου ή αλλιώς 'text mining' αναφέρεται στην διαδικασία ανάλυσης, ανακάλυψης μοτίβων, εξαγωγής συμπερασμάτων και γνώσης από αδόμητα δεδομένα, αποθηκευμένα σε φυσική γλώσσα. Ο κλάδος του text mining έχει σημειώσει τεράστια άνοδο τα τελευταία χρόνια καθώς έχει γίνει αντιληπτό πως αν τα διαθέσιμα δεδομένα αξιοποιηθούν επιτυχώς, υπάρχει θετικός αντίκτυπος στην πορεία της εταιρείας ή του οργανισμού. Ενδεικτική είναι μια έρευνα[6] δημοσιευμένη από το 'McKinsey Global Institute' η οποία υποστηρίζει πως οι ευρωπαϊκές κυβερνήσεις θα μπορούσαν να εξοικονομούν 100 δισεκατομμύρια ευρώ τον χρόνο αν αξιοποιούσαν διαθέσιμα δεδομένα πιο αποτελεσματικά. Φαίνεται λοιπόν, πως είναι πολύ σημαντικό για επιχειρήσεις και οργανισμούς να αναπτυχθούν τεχνολογίες οι οποίες αναλύουν και αξιοποιούν δεδομένα αποτελεσματικά σε μικρό χρονικό διάστημα.

Όμως, κατά την ανάλυση δεδομένων γραμμένων σε φυσική γλώσσα παρουσιάζονται και πολλές προκλήσεις που πηγάζουν κυρίως από την φύση των δεδομένων. Κάποια από τα εμπόδια που πρέπει να ξεπεραστούν κατά την διαδικασία του text mining είναι τα εξής:

- Η φυσική γλώσσα είναι περίπλοκη, πολλές φορές αδόμητη με πολύ θόρυβο, γεμάτη ασάφειες και υπαινιγμούς.
- Είναι αναπόφευκτη η ύπαρξη ειρωνείας, μεταφορικού λόγου, λαϊκών εκφράσεων, διαφορετικών διαλέκτων καθώς και ομόγραφων λέξεων σε κείμενα φυσικής γλώσσας.
- Τα διαθέσιμα δεδομένα είναι γραμμένα σε πολλές διαφορετικές γλώσσες. Μία αναφορά[7] από το statista[8] αναφέρει πώς μόνο το 25.9% των χρηστών του διαδικτύου χρησιμοποιεί τα αγγλικά, κάτι που σημαίνει πως τα 3/4 της διαθέσιμης πληροφορίας στο διαδίκτυο είναι σε άλλη γλώσσα.
- Οι πηγές των δεδομένων είναι πάρα πολλές και τα δεδομένα αποθηκεύονται σε πολλές διαφορετικές μορφές.
- Η ανανέωση των δεδομένων είναι ταχύτατη, αδιάκοπη και τεράστια σε όγκο πληροφορίας. Για παράδειγμα, στο twitter[9] από το 2013 δημοσιεύονται καθημερινά κατά μέσο όρο 500 εκατομμύρια νέα 'tweets'.

Είναι λοιπόν ξεκάθαρο, πως η εξόρυξη γνώσης από δεδομένα σε φυσική γλώσσα είναι μια

πρόκληση αλλά ταυτόχρονα και μια διαδικασία η οποία είναι αναγκαία σε επιχειρήσεις για να ληφθούν κομβικής σημασίας αποφάσεις για το μέλλον.

Λαμβάνοντας υπόψη τον όγκο το διαθέσιμων δεδομένων σήμερα αλλά και αυτών που παράγονται καθημερινά, είναι πολύ κοστοβόρο για κάποια διαδικασία να επεξεργαστούν όλα τα διαθέσιμα δεδομένα που την αφορούν, ή πολλές φορές και μη εφικτό. Υπάρχει λοιπόν η ανάγκη για ανάπτυξη τεχνολογιών, οι οποίες θα εξάγουν ένα αποτέλεσμα για κάποια διεργασία, χωρίς να επεξεργάζονται το σύνολο το διαθέσιμων δεδομένων αλλά ένα μικρό υποσύνολο αυτών που θα περιέχει την πληροφορία που απαιτείται.

Α΄.4 Εφαρμογές που βασίζονται σε δεδομένα φυσικής γλώσσας

Ένα πολύ σημαντικό χαρακτηριστικό της φυσικής γλώσσας είναι ότι εμπεριέχει έμμεση πληροφορία, καθώς μέσα από ένα κείμενο ή μία φράση μπορείς να εξάγεις συμπεράσματα τα οποία δεν αναφέρονται σε αυτό, σε αντίθεση με τα αριθμητικά δεδομένα τα οποία απλά αναπαριστούν μία ποσότητα. Για παράδειγμα, η φράση 'Αυτό το προϊόν δεν αξίζει τα λεφτά του!' υποδηλώνει και την δυσαρέσκεια του προσώπου που την αναφέρει, πέρα από το προφανές νόημα. Αυτό το ιδιαίτερο στοιχείο κατατάσει τα δεδομένα φυσικής γλώσσας σε μία βασική πηγή γνώσης, και γι΄ αυτό συνεχώς ερευνώνται νέες τεχνικές για να αξιοποιηθούν.

Κάποια παραδείγματα εφαρμογών που εκμεταλλεύονται και αναλύουν δεδομένα φυσικής γλώσσας είναι:

- Οι αυτόματες μεταφράσεις κειμένων σε διάφορες γλώσσες, όπου εισάγοντας ένα κείμενο σε κάποια γλώσσα, η εφαρμογή την μεταφράζει αυτόματα σε κάποια άλλη [10] .
- Οι ανιχνευτές ανεπιθύμητων και κακόβουλων μηνυμάτων. Πρόκειται για εφαρμογές οι οποίες εντοπίζουν μηνύματα ή emails που θεωρούνται μη αξιόπιστα. Συνήθως η κατηγοριοποίηση βασίζεται στο περιεχόμενο αλλά και την πηγή του μηνύματος, και υλοποιείται με τεχνικές μηχανικής μάθησης. Τέτοιες τεχνολογίες συναντώνται συχνά στα κοινωνικά δίκτυα και τις εφαρμογές ηλεκτρονικού ταχυδρομείου[11][12].
- Οι ψηφιακοί βοηθοί και τα Chatbots. Είναι εφαρμογές οι οποίες απαντούν σε ερωτήσεις χρηστών που παρατίθενται σε φυσική γλώσσα, αλλά υπάρχουν και εναλλακτικές οι οποίες απαντούν σε φωνητικές εντολές. Η πιο απλή υλοποίηση απαντάει στις ερωτήσεις βάσει μιας βάσης δεδομένων που αποθηκεύει ζευγάρια ερωτήσεων-απαντήσεων, ενώ υπάρχουν και πιο περίπλοκες υλοποιήσεις που κάνουν χρήση μηχανικής μάθησης[13][14].
- Οι εφαρμογές εξόρυξης γνώμης ή ανάλυσης συναισθήματος(Sentiment Analysis), όπου στόχος είναι να αναλυθεί η γνώμη κάποιου ατόμου για ένα προϊόν, μία υπηρεσία ή ένα θέμα βάσει κάποιου σχολίου που έχει κάνει [15].
- Οι εφαρμογές αυτόματου συνοψισμού μεγάλων κειμένων, οι οποίες λαμβάνουν σαν είσοδο ένα κείμενο και παράγουν αυτόματα μια περίληψη του όπου αναφέρονται τα πιο σημαντικά στοιχεία του[16].

Α΄.5 Επεξεργασία φυσικής γλώσσας

Ο κλάδος της επεξεργασίας φυσικής γλώσσας στην επιστήμη των υπολογιστών αφορά όλες τις ενέργειες που πραγματοποιούνται ώστε μία υπολογιστική μηχανή να επεξεργάζεται τον προφορικό και γραπτό λόγο σαν να ήταν άνθρωπος, και να μπορεί να εξάγει συμπεράσματα σαν αυτά του ανθρώπινου εγκεφάλου. Σκοπός είναι να δημιουργηθούν αποτελεσματικές μέθοδοι που θα βοηθούν έναν υπολογιστή να αντιλαμβάνεται το άμεσο και έμμεσο νόημα ενός κειμένου, χωρίς να απαιτούνται επιπλέον πληροφορίες ή ενέργειες. Το όφελος από αυτή την διαδικασία είναι το μέγεθος των κειμένων και των πληροφοριών που μπορούν να αναλυθούν και να αντληθούν αντίστοιχα μέσα σε μικρό χρονικό διάστημα, κάτι που θα ήταν αδύνατο για έναν ανθρώπινο εγκέφαλο.

Η επεξεργασία φυσικής γλώσσας για να είναι ολοκληρωμένη θα μπορούσε να χωριστεί σε τρία στάδια, την προεπεξεργασία, την εφαρμογή της μεθόδου και την εξαγωγή αποτελέσματος. Στο στάδιο της προεπεξεργασίας το κείμενο διαμορφώνεται κατάλληλα ώστε να είναι πιο αποτελεσματική η επεξεργασία του, για παράδειγμα καθαρίζεται από θόρυβο και αχρείαστες λέξεις. Στην συνέχεια στο στάδιο της εφαρμογής της μεθόδου εφαρμόζεται σε αυτό η επιλεγμένη τεχνική ώστε να εξαχθεί κάποιο αποτέλεσμα, για παράδειγμα κατηγοριοποιείται το κείμενο βάσει του περιεχομένου του, και στο τελευταίο στάδιο εξάγεται το αποτέλεσμα της επεξεργασίας. Ένα ολοκληρωμένο παράδειγμα είναι η διαδικασία της περίληψης ενός κειμένου, όπου πρώτα το κείμενο τροποποιείται για να είναι δυνατή η περίληψη του (εντοπίζονται οι λέξεις κλειδιά, λέξεις τροποποιούνται), στην συνέχεια το κείμενο συμπυκνώνεται βάσει των σημαντικότερων πληροφοριών και τέλος αποδίδεται η περίληψη του.

Για να εφαρμοστεί η επεξεργασία φυσικής γλώσσας υπάρχουν πλέον πολλές τεχνικές και εργαλεία που απλοποιούν την διαδικασία, και η πρόκληση είναι να συνδυαστούν κατάλληλα ανάλογα με την περίπτωση. Κάποιες από τις κυριότερες βιβλιοθήκες για επεξεργασία φυσικής γλώσσας είναι η NLTK, όπου παρέχει δυνατότητες για επεξεργασία αλλά και αλγορίθμους ταξινόμησης, η scikit–learn η οποία βοηθάει πολύ στην δημιουργία μοντέλων πρόβλεψης, η Gensim όπου ενδείκνυται για μετατροπές κειμένων σε διανύσματα και υπολογισμό νοηματικής ομοιότητας μεταξύ κειμένων και η spaCy η οποία παρέχει όλα τα βασικά εργαλεία για επεξεργασία φυσικής γλώσσας και είναι πολύ αποδοτική για μεγάλου μεγέθους δεδομένα.

Α΄.6 Περίληψη προτεινόμενης μεθόδου

Στην παρούσα εργασία προτείνεται μια εναλλακτική λύση για εργασίες μεγάλου όγκου δεδομένων στις οποίες κάποιος αλγόριθμος εφαρμόζεται σε δεδομένα κειμένου και δέχεται σαν είσοδο ένα μόνο κείμενο. Η υλοποίηση βασίζεται στην ιδέα[17] πως είναι δυνατόν να προβλεφθεί το αποτέλεσμα ενός αλγορίθμου για ένα μεγάλο σύνολο δεδομένων, αν ο αλγόριθμος εφαρμοστεί σε ένα μικρό υποσύνολο αυτού. Σκοπός της εργασίας είναι να προτείνει μια μεθοδολογία η οποία θα μπορούσε υπο προϋποθέσεις να μειώσει την χρονική πολυπλοκότητα ενός χρονοβόρου τελεστή, κάνοντας μία πρόβλεψη για την έξοδο του αν εισάγονταν σε αυτόν συγκεκριμένα δεδομένα. Θεωρώντας πως θέλουμε να εφαρμόσουμε έναν πολύπλοκο αλγόριθμο σε δεδομένα κειμένου, επιδιώκεται να μειωθεί ο απαιτούμενος χρόνος εισάγοντας σε αυτόν ένα

πολύ μικρό ποσοστό των δεδομένων και προβλέποντας το αποτέλεσμά του για τα υπόλοιπα.

Η προτεινόμενη μέθοδος που υλοποιήθηκε και αναλύεται στο μέρος *Β΄* είναι η εξής: Έστω ότι έχουμε ένα σύνολο δεδομένων κειμένου D, N εγγραφών, στο οποίο θέλουμε να εφαρμόσουμε έναν αλγόριθμο σε κάθε μία εγγραφή ξεχωριστά. Υπολογίζοντας την ομοιότητα ή 'Similarity' όλων των N εγγραφών μεταξύ τους, όπου 'Similarity' ένας αριθμός $x \in [0, 1]$, κατασκευάζουμε έναν πίνακα *(N*N)* 'Similarity Matrix' με τα Similarities των εγγραφών ανά μεταξύ τους. Στην συνέχεια, εφαρμόζουμε τον αλγόριθμο σε ένα υποσύνολο Di του D, $Di \subseteq D$, και με βάση την έξοδο του αλγορίθμου για τις εγγραφές του Di και τον Similarity Matrix, προβλέπουμε την έξοδο του αλγορίθμου για το σύνολο Dj, $Dj \subseteq D$ και $Dj \cup Di = D$.

Κεφάλαιο Β΄

Κύριο Μέρος

Β΄.1 Προεπεξεργασία Φυσικής Γλώσσας

Για να είναι πιο ξεκάθαρη και εύχρηστη η πληροφορία που εμπεριέχεται σε κείμενα και δεδομένα φυσικής γλώσσας, είναι απαραίτητη η προεπεξεργασία τους. Το στάδιο της προεπεξεργασίας είναι το πρώτο βήμα κατά την διαδικασία του 'text mining' και εχει ως κύριο στόχο την προετοιμασία του κειμένου για την ανάλυση. Ο όρος προεπεξεργασία αναφέρεται στις τεχνικές που έχουν ως στόχο την αφαίρεση όλων των στοιχείων τα οποία δεν επηρεάζουν το κείμενο νοηματικά, καθώς και την τροποποίηση του ώστε να αφαιρεθεί ο θόρυβος και να είναι πιο ξεκάθαρη η πληροφορία. Ένα παράδειγμα στοιχείου που δεν επηρεάζει το νόημα του κειμένου συναντάται συνήθως σε δημοσιεύσεις στα μέσα κοινωνικής δικτύωσης, οι οποίες πολύ συχνά περιέχουν υπερσυνδέσμους. Οι υπερσύνδεσμοι σαν ακολουθία χαρακτήρων σε φυσική γλώσσα δεν έχουν κάποιο νόημα και κατα συνέπεια η αφαίρεση τους δεν αλλοιώνει τη γενικότερη πληροφορία του κειμένου . Οι τεχνικές προεπεξεργασίας ποικίλλουν ανάλογα με τον σκοπό της ανάλυσης, και μπορούν να εξατομικευτούν σύμφωνα με την εκάστοτε περίπτωση.

Β΄.1.1 Αφαίρεση κενών λέξεων και μετατροπή γραμμάτων σε πεζά

Μια από τις βασικές τεχνικές προεπεξεργασίας κειμένου είναι η αφαίρεση λέξεων οι οποίες δεν έχουν κάποια επίδραση στο γενικότερο νόημα του κειμένου. Είναι οι λεγόμενες 'κενές λέξεις' ή στα αγγλικά 'stop words'. Πρόκειται για λέξεις οι οποίες συμβάλλουν κυρίως στην σωστή γραμματική και συντακτική δομή του κειμένου και όχι τόσο στην πληροφορία που εμπεριέχεται στο κείμενο. Παραδείγματα τέτοιών λέξεων είναι άρθρα(ο, τους, της, των), σύνδεσμοι(και, ή, ώστε, όταν, ενώ), επίρρηματα, αντωνυμίες(μου, σου, που, οι οποίοι, οτι, όσες). Η πιο συνήθης πρακτική που χρησιμοποιείται για την αφαίρεση τέτοιων λέξεων είναι η χρήση βιβλιοθηκών[18], που περιέχουν πολλές κενές λέξεις σε διάφορες γλώσσες και τις αφαιρούν αυτόματα. Επίσης, παράλληλα με την αφαίρεση κενών λέξεων μπορούν να αφαιρεθούν και τα σημεία στίξης, καθώς δεν προσφέρουν κάποια πληροφορία ούτε και επηρεάζουν το νόημα του κειμένου.

Μια άλλη πολύ σημαντική τεχνική είναι η μετατροπή όλων των γραμμάτων σε πεζά (Lowercasing). Στην γλώσσα μηχανής, η κεφαλαία και η πεζή μορφή ενός γράμματος είναι δύο διαφορετικοι χαρακτήρες. Οπότε, σε περίπτωση που δύο λέξεις είναι ίδιες αλλά η μία έχει κάποιο γράμμα κεφαλαίο, ο υπολογιστής τις διαβάζει σαν δύο διαφορετικές με αποτέλεσμα να δυσχεραίνεται η διαδικασία της εξαγωγής γνώσης.

Β΄.1.2 Stemming και Lemmatization

Οι τεχνικές Stemming και Lemmatization αντικαθιστούν τις λέξεις του κειμένου με τις αντίστοιχες λέξεις που έχουν σαν ρίζα αλλά με διαφορετικές προσεγγίσεις. Ένα παράδειγμα διαφορετικών λέξεων που έχουν την ίδια ρίζα είναι το 'Είδα' και το 'Έβλεπα', όπου και των δύο η ρίζα είναι το 'Βλέπω', νοηματικά είναι συγγενικές αλλά ο υπολογιστής της διαβάζει σαν δύο τελείως διαφορετικές λέξεις. Ο σκοπός είναι να αντιμετωπίζονται από το εκάστοτε πρόγραμμα σαν ίδιες λέξεις όσες λέξεις έχουν κοινή ρίζα, γι΄ αυτό και αντικαθίστανται από την ίδια λέξη.

Η τεχνική Stemming για να αντικαταστήσει τις λέξεις με την ρίζα τους αφαιρεί, ανάλογα με την περιπτώση, την κατάληξη της λέξης. Αυτή η τεχνική μπορεί αρκετές φορές να μην παράγει την πραγματική ρίζα της λέξης, καθώς απλά αφαιρεί τα τελευταία γράμματα, αλλά είναι μια προσέγγιση η οποία θα μπορούσε να συνεισφέρει στην διαδικασία της προεπεξεργασίας κάποιου κειμένου. Ένας από τους πιο συνηθισμένους αλγορίθμους που χρησιμοποιείται για να εφαρμοστεί το Stemming είναι ο αλγόριθμος του Porter[19].

Το Lemmatization από την άλλη, μετατρέπει τις λέξεις ενός κειμένου στην ρίζα τους αλλά όχι απλά κόβοντας την κατάληξη. Βασίζεται είτε σε κανόνες που δίνονται σαν παράμετροι, είτε σε λεξικές βάσεις δεδομένων που περιέχουν τις αντιστοιχίες λέξεων με την ρίζα τους. Μια μεγάλη διαφορά των δύο τεχνικών είναι ότι το Lemmatization μπορεί και κατηγοριοποιεί νοηματικά συγγενικές λέξεις στην ίδια ρίζα ακόμα και αν δεν προέρχονται από την ίδια λέξη (π.χ ἀμάξι΄ και ἀυτοκίνητο΄) , ενώ αντίθετα το Stemming βασίζεται αποκλειστικά και μόνο στον τρόπο που γράφεται μια λέξη. Μια πολύ γνωστή βάση δεδομένων λέξεων που χρησιμοποιείται για την εφαρμογή του Lemmatization είναι το WordNet[20].

Β΄.1.3 Tokenization

Μια άλλη σημαντική τεχνική που εφαρμόζεται στα δεδομένα κατά το στάδιο της προεπεξεργασίας είναι το 'Tokenization'. Πρόκειται για τον διαχωρισμό του κειμένου σε λέξεις ή προτάσεις, ανάλογα με το πρόβλημα. Το κείμενο μετατρέπεται σε μια λίστα όπου κάθε στοιχείο της είναι μία λέξη ή πρόταση του κειμένου. Αυτή η τεχνική είναι πολύ βασική για εργασίες όπου θέλουμε να συγκρίνουμε δύο κείμενα, να δούμε τις πιο συχνές λέξεις σε ένα κείμενο ή να δούμε πόση βαρύτητα έχει μια λέξη σε ένα κείμενο.

Β΄.1.4 Vectorization

Η μέθοδος 'Vectorization', χρησιμοποιείται για την μετατροπή λέξεων ή κειμένων από φυσική γλώσσα σε αριθμητικά διανύσματα. Υπάρχουν πολλές διαφορετικές τεχνικές για την εφαρμογή

της μεθόδου και χρησιμοποιούνται ανάλογα με τον σκοπό της εργασίας.

Η πιο απλή τεχνική για *Vectorization* είναι η *'Bag of Words'*, όπου σε όλα τα κείμενα των δεδομένων εισόδου εφαρμόζεται το Tokenization, δημιουργείται ένα λεξικό με όλες τις μοναδικές λέξεις των κειμένων και η κάθε εγγραφή των δεδομένων εισόδου μετατρέπεται σε διάνυσμα, μήκους ίσου με το μέγεθος του λεξικού, και τιμές ανάλογα με την συχνότητα εμφάνισης μιας λέξης του λεξικού στην εγγραφή. Με την ίδια λογική, το λεξικό θα μπορούσε να είναι συνδυασμοί n λέξεων και οι τιμές του διανύσματος οι εμφανίσεις τους.
Υπάρχουν και πιο περίπλοκες τεχνικές για Vectorization, όπως η TF-IDF[21] και GloVe[22], όπου το διάνυσμα των λέξεων ή προτάσεων δημιουργείται δίνοντας έμφαση στην βαρύτητα των λέξεων καθώς και στην σημασία τους.

Το στάδιο της προεπεξεργασίας είναι πολύ σημαντικό για εργασίες που θέλουν να επεξεργαστούν και να αξιοποιήσουν δεδομένα κειμένου, καθώς τα δεδομένα φυσικής γλώσσας έχουν πολλές ιδιαιτερότητες και απαιτείται η κατάλληλη προεργασία για να αξιοποιηθούν κατάλληλα. Οι παραπάνω τεχνικές είναι κάποιες από τις πιο βασικές για την προεπεξεργασία κειμένου και εφαρμόζονται κατάλληλα, ανάλογα με το ζητούμενο αλλά και την φύση των δεδομένων. Υπάρχουν πολλές παραλλαγές τους για διαφορετικά σενάρια και το πώς θα χρησιμοποιηθούν παραμένει στην ευχέρεια του εκάστοτε αναλυτή.

Β΄.2 Υπολογισμός Ομοιότητας Μεταξύ των Δεδομένων

Η ομοιότητα μεταξύ δύο κειμένων είναι μία πρόκληση η οποία απασχολεί πολλούς ερευνητές εδώ και αρκετά χρόνια. Το κύριο πρόβλημα στην ομοιότητα μεταξύ δύο κειμένων είναι το ποιά στοιχεία τους τα χαρακτηρίζουν όμοια. Οι απαντήσεις σε αυτό είναι πολλές, και πάντα εξαρτώνται από τον σκοπό της κάθε εργασίας.

Μία από τις πρώτες προσεγγίσεις για τον υπολογισμό της ομοιότητας μεταξύ δύο προτάσεων γραμμένων σε φυσική γλώσσα έγινε το 1977 από τον George H. Stalker[23]. Η προτεινόμενη μέθοδος έχει 3 διαφορετικές υλοποίησεις, δίνοντας βάση σε διαφορετικά στοιχεία των δύο προτάσεων. Η πρώτη βασίζεται στο πόσες λέξεις έχουν κοινές οι δύο προτάσεις, η δεύτερη στο πόσες ακολουθίες χαρακτήρων σταθερού μήκους έχουν κοινές, και η τρίτη μετράει ασυνέχειες κατά το ταίριασμα ακολουθιών χαρακτήρων. Απο τότε και μέχρι σήμερα έχουν προταθεί πάρα πολλές τεχνικές για τον υπολογισμό της ομοιότητας μεταξύ δύο προτάσεων ή κειμένων, δίνοντας βάση σε διαφορετικά χαρακτηριστικά και πολλές φορές κάνοντας περίπλοκους συνδυασμούς [24], [25], [26].

Για την υλοποίηση της μεθόδου που προτείνεται στην παρούσα εργασία,εφαρμόστηκαν και δοκιμάστηκαν 3 διαφορετικές τεχνικές υπολογισμού της ομοιότητας μεταξύ 2 κειμένων, όπου η κάθε μία έχει διαφορετική προσέγγιση για την ομοιότητα. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι εξής τεχνικές:

- Euclidean Distance

- Cosine Similarity
- Word Movers Distance

Για να υπολογιστεί η απόσταση ή η ομοιότητα 2 κειμένων, είναι απαραίτητες κάποιες ενέργειες προεπεξεργασίας. Οι περισσότεροι αλγόριθμοι υπολογισμού απόστασης ή ομοιότητας, που βασίζονται στις λέξεις των δύο κειμένων, προϋποθέτουν τα κείμενα να αναπαρασταθούν σε διανύσματα σε έναν διανυσματικό χώρο. Πρώτα, εφαρμόζεται στο κείμενο η τεχνική του Tokenization, αφού έχει προηγηθεί η βασική προεπεξεργασία, και στην συνέχεια εφαρμόζεται η μέθοδος Vectorization ώστε τα κείμενα να μετατραπούν σε διανύσματα. Έτσι,με βάση τα διανύσματα και το μέτρο που έχουμε επιλέξει, υπολογίζεται η ομοιότητα των δύο κειμένων.

Β΄.2.1 Euclidean Distance

Η Ευκλείδεια απόσταση $||x\text{-}y||_2$ δύο διανυσμάτων $x = (x_1, x_2, \ldots, x_n)$ και $y = (y_1, y_2, \ldots, y_n)$ υπολογίζεται βάσει τον τύπο:

$$||x - y||_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

και είναι μία αριθμητική τιμή $d \in [0, 1]$, μετά από κανονικοποίηση, όπου αν $d = 0$ οι προτάσεις είναι πανομοιότυπες και αν $d = 1$ οι προτάσεις είναι τελείως ανόμοιες.

Β΄.2.2 Cosine Similarity

Το Cosine Similarity υπολογίζεται από την γωνία που σχηματίζουν δύο διανύσματα $x = (x_1, x_2, \ldots, x_n)$ και $y = (y_1, y_2, \ldots, y_n)$ σε ένα διανυσματικό χώρο. Ο τύπος είναι:

$$\text{cosine similarity} = \cos(\theta) = \frac{x \cdot y}{\| x \| \cdot \| y \|}$$

Β΄.2.3 Word Movers Distance

Το μέτρο Word Movers Distance(WMD)[27] υπολογίζει την απόσταση μεταξύ δύο κειμένων δίνοντας έμφαση και στην νοηματική συσχέτιση τους, όχι μόνο στις κοινές τους λέξεις. Η νοηματική συσχέτιση των δύο κειμένων βασίζεται σε Word Embeddings, τα οποία είναι προϋπάρχουσες αναπαραστάσεις εκατομμυρίων λέξεων σε έναν πολυδιάστατο διανυσματικό χώρο, όπου τα διανύσματα των λέξεων που χρησιμοποιούνται με τον ίδιο τρόπο σε εκφράσεις ή έχουν παρόμοια ετυμολογία βρίσκονται σε κοντινή απόσταση. Η πιο ευρέως γνωστή υλοποίηση των Word Embeddings, που χρησιμοποιείται και στο WMD, είναι το Word2Vec[28].
Με βάση λοιπόν τα Word Embeddings, ορίζεται η απόσταση των δύο κειμένων από τα εξής 3 μέρη:

- Το κάθε κείμενο αναπαρίσταται σαν ένα πολυδιάστατο διάνυσμα $d = [d_1, d_2, \ldots, d_n]^T$,όπου $d_i = \frac{c_i}{\sum_{j=1}^{n} c_j}$ και c_i ={οι εμφανίσεις της λέξης i σε ένα κείμενο}.
- Σαν 'κόστος ταξιδιού' $c(i, j)$ μίας λέξης i σε μία λέξη j από διαφορετικό κείμενο, ορίζεται η Ευκλείδεια απόσταση τους στον διανυσματικό χώρο των 'Word Embeddings'. Δηλαδή, $c(i, j) = \| x_i - x_j \|_2$
- Υποθέτοντας πως έχουμε δύο κείμενα **d** και **d'**, μετατρέποντας κάθε λέξη i του **d** σε μία λέξη j του **d'**, ορίζουμε σαν $\mathbf{T}$ έναν πίνακα $n \times n$ όπου κάθε $\mathbf{T}_{ij} \geq 0$ αναπαριστά το *πόσο i πρέπει να μετακινηθεί για να φτάσει το j.*

Με αυτά υπόψιν, σαν WMD δύο κειμένων **d** και **d'** ορίζεται:

$$distance = \min_{T \geq 0} \sum_{i,j=0}^{n} \mathbf{T}_{i,j} c(i, j)$$

με:

$$\sum_{j=1}^{n} \mathbf{T}_{i,j} = d_i \qquad \forall i \in \{1, \ldots, n\}$$

και:

$$\sum_{i=1}^{n} \mathbf{T}_{i,j} = d'_j \qquad \forall j \in \{1, \ldots, n\}$$

Β΄.2.4 Εργαλεία που Χρησιμοποιήθηκαν

Για την υλοποίηση και εφαρμογή όλων των προαναφερθέντων τεχνικών χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, καθώς και διάφορες βιβλιοθήκες για επεξεργασία φυσικής γλώσσας. Για τις ανάγκες υλοποίησης της κάθε τεχνικής υπολογισμού ομοιότητας μεταξύ δύο κειμένων, χρησιμοποιήθηκαν και διαφορετικοί συνδυασμοί προεπεξεργασίας κειμένου.

Όσον αφορά την υλοποίηση της Ευκλείδειας απόστασης, έχοντας ένα σύνολο δεδομένων που αποτελείται από n διαφορετικά κείμενα, εφαρμόζεται η μέθοδος του Tokenization,αφού από τα κείμενα έχουν αφαιρεθεί ειδικοί χαρακτήρες και σημεία στίξης. Στην συνέχεια δημιουργείται ένα λεξικό με το σύνολο των λέξεων από όλα τα κείμενα, και αυτά μετατρέπονται σε πολυδιάστατα διανύσματα σύμφωνα με το αν μια λέξη του λεξικού υπάρχει στο κείμενο. Μετατρέποντας το κάθε κείμενο σε διάνυσμα, υπολογίζεται η Ευκλείδεια απόσταση όλων των κειμένων μεταξύ τους και τα αποτελέσματα αποθηκεύονται σε έναν πίνακα $SimilarityMatrix^{n \times n}$. Με την ίδια ακριβώς διαδικασία υπολογίζεται και το Cosine Similarity, με τα αποτελέσματα επίσης να αποθηκεύονται σε έναν πίνακα $SimilarityMatrix^{n \times n}$

Κατά την υλοποίηση του WMD, για Word Embeddings χρησιμοποιήθηκε το προκατασκευασμένο μοντέλο word2vec-google-news-300 από την βιβλιοθήκη gensim[29]. Πρώτα αφαιρούνται από τα n κείμενα όλα τα σημεία στίξης και οι ειδικοί χαρακτήρες, στην συνέχεια υπολογίζεται το WMD μεταξύ όλων των προτάσεων και τα αποτελέσματα αποθηκεύονται σε έναν πίνακα $SimilarityMatrix^{n \times n}$.

Β΄.3 Τελεστές που Χρησιμοποιήθηκαν

Υπάρχουν πολλοί διαφορετικοί τελεστές που εφαρμόζονται σε δεδομένα κειμένου ώστε να εξαχθεί κάποια γνώση από αυτά, όπως αλγόριθμοι ταξινόμησης, ομαδοποίησης, ανάλυσης συναισθήματος, αναζήτησης λέξεων, εύρεσης συσχετίσεων. Η κύρια πρόκληση της παρούσας εργασίας, είναι να βρεθεί κάποιος τελεστής ο οποίος αν δεχθεί σαν είσοδο ένα κείμενο το οποίο έχει μικρή απόσταση με κάποιο άλλο, αυτά να έχουν και την ίδια έξοδο στον τελεστή με μεγάλη πιθανότητα.

Για να γίνει πιο σαφές το πρόβλημα, ας υποθέσουμε πως θέλουμε να εφαρμόσουμε έναν αλγόριθμο ανάλυσης συναισθήματος *(Sentiment Analysis)* σε ένα μεγάλο αριθμό δημοσιεύσεων από το διαδίκτυο. Για να μπορέσουμε να προβλέψουμε το αν μια δημοσίευση υποδηλώνει ότι ο συντάκτης της έχει θετική ή αρνητική διάθεση, θα πρέπει κατά γενικό κανόνα οι συντάκτες δημοσιεύσεων με μεγάλη ομοιότητα να έχουν και ίδια διάθεση. Είναι πολύ σημαντικό αυτό το κομμάτι καθώς πάνω στην ομοιότητα των δεδομένων εισόδου βασίζεται και η πρόβλεψη της εξόδου. Θεωρώντας 2 δημοσιεύσεις, *'Σήμερα είναι μια υπέροχη μέρα'* και *'Σήμερα είναι μια υπέροχη μέρα, αλλά δυστυχώς είμαι άρρωστος'*, μπορούμε εύκολα να υποθέσουμε πως ο συντάκτης της πρώτης δημοσίευσης έχει θετική διάθεση ενώ της δεύτερης αρνητική. Με βάση όμως τα μέτρα ομοιότητας που αναλύθηκαν παραπάνω, οι δύο προτάσεις έχουν μικρή απόσταση ή αντίστοιχα είναι αρκετά όμοιες. Αυτή λοιπόν είναι η μεγάλη πρόκληση στην παρούσα εργασία, να βρεθούν αλγόριθμοι οι οποίοι έχουν ίδια έξοδο για κείμενα με μεγάλη ομοιότητα.

Μετά από πολλές δοκιμές, ένας αλγόριθμος που φάνηκε να πληροί την παραπάνω βασική προϋπόθεση, είναι αυτός της αναγνώρισης γλώσσας. Είναι λογικό, πως κείμενα γραμμένα σε διαφορετική γλώσσα δεν παρουσιάζουν πολλές ομοιότητες όσον αφορά τις λέξεις που χρησιμοποιούνται, που είναι βασικό στοιχείο του υπολογισμού της απόστασης, άρα θεωρητικά θα έχουν μεγάλη απόσταση και αντίστοιχα κείμενα στην ίδια γλώσσα μικρότερη. Πιθανώς, να ήταν πολύ χρήσιμη για αναλυτικές εργασίες μεγάλου όγκου δεδομένων η πρόβλεψη της γλώσσας πολλών κειμένων, με μεγάλη ακρίβεια και χωρίς να χρειάζεται να εφαρμοστεί σε όλα κάποιος αλγόριθμος αναγνώρισης γλώσσας. Πρέπει επίσης να αναφερθεί, πως είναι πίο εποικοδομητική μία τέτοια πρόβλεψη αν εφαρμοστεί, και είναι αποτελεσματική, για γλώσσες οι οποίες μοιράζονται το ίδιο αλφάβητο, καθώς σε τέτοιες περιπτώσεις υπάρχει περίπτωση να περιλαμβάνουν κοινές λέξεις ή ακολουθίες χαρακτήρων.

Σαν αλγόριθμος αναγνώρισης γλώσσας, χρησιμοποιήθηκε η βιβλιοθήκη langdetect[30] η οποία αναγνωρίζει κείμενα σε 53 διαφορετικές γλώσσες.

Β΄.4 Δεδομένα που Συλλέχθηκαν

Για την υλοποίηση της μεθόδου, επιλέχθηκε η εφαρμογή της σε δεδομένα από 5 λατινογενής γλώσσες, με διαφορετικές θεματολογίες και πηγές. Τα σύνολα δεδομένων που συλλέχθηκαν για τις ανάγκες της εργασίας είναι από κείμενα σε:

- Αγγλικά

- Γερμανικά
- Ισπανικά
- Ιταλικά
- Γαλλικά

Επιλέχθηκαν κείμενα με σχετικά μικρό μέγεθος, ώστε να μην είναι πολύ ξεκάθαρη η ομοιότητα μεταξύ κειμένων σε ίδια γλώσσα. Οι τύποι των κειμένων διαφέρουν, χωρίς να υπάρχει κάποιο κριτήριο επιλογής, πέρα από την γλώσσα που είναι γραμμένα, για να υπάρχει όσο το δυνατόν μεγαλύτερη τυχαιότητα.

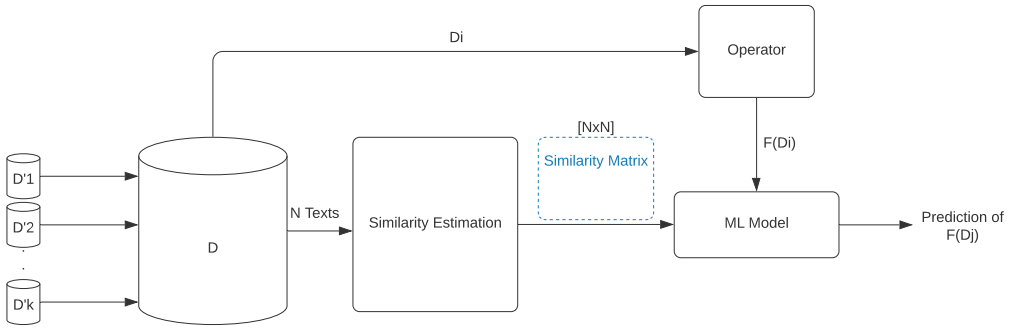
Β΄.4.1 Πηγές δεδομένων

- Τα αγγλικά κείμενα που χρησιμοποιήθηκαν είναι προτάσεις μικρού μεγέθους από ένα σύνολο δεδομένων που έχει δημοσιευθεί στο kaggle[31]. Η επιλογή από το σύνολο δεδομένων έγινε τυχαία.
- Τα κείμενα στα γερμανικά που επιλέχθηκαν, είναι προτάσεις από μία πληθώρα άρθρων απο το Wikipedia με διάφορες θεματολογίες. [32]
- Τα Ισπανικά κείμενα που επιλέχθηκαν, είναι μέρος μιας συλλογής προτάσεων που συλλέχθηκαν από το Wikipedia. [33]
- Όσον αφορά τα δεδομένα στα ιταλικά,χρησιμοποιήθηκε κομμάτι από μια συλλογή Tweets χρηστών από διαφορετικές περιοχές της Ιταλίας, με διαφορετικές διαλέκτους. [34]
- Για τα Γαλλικά κείμενα, έγινε χρήση ενός συνόλου τυχαίων καθημερινών προτάσεων μεταφρασμένων από τα αγγλικά. [35]

Ο συνολικός αριθμός των διαφορετικών κειμένων που συλλέχθηκαν ανέρχεται σε εκατομμύρια. Για να μπορέσει όμως να δοκιμαστεί η μέθοδος επιλέχθηκαν τυχαία 2.000 εγγραφές από κάθε ένα από τα προαναφερθέντα σύνολα δεδομένων, όποτε το τελικό σύνολο αποτελείται από 10.000 διαφορετικά κείμενα.

Β΄.5 Μέθοδος πρόβλεψης

Όπως έχει προαναφερθεί, στόχος της παρούσας μεθόδου είναι να προβλεφθεί η έξοδος ενός τελεστή για δεδομένα κειμένου ενός συνόλου δεδομένων D, χωρίς να εφαρμοστεί ο τελεστής σε όλο το σύνολο. Βάσει λοιπόν τον *Similarity Matrix* και το αποτέλεσμα $(F(D_i))$ του τελεστή για ένα υποσύνολο D_i του συνόλου των δεδομένων D, με $D_i \subseteq D$, προβλέπουμε το αποτέλεσμα $(F(D_j))$ του τελεστή για το υποσύνολο D_j, με $D_i \cup D_j = D$. Στο σχήμα Β΄1 δίνεται μία αναπαράσταση των διεργασιών της μεθοδολογίας.



Σχήμα Β΄.1: *Ροή εργασιών μεθοδολογίας*

Υπάρχουν διάφορες τεχνικές που θα μπορούσαν να πραγματοποιήσουν μια τέτοια πρόβλεψη, όπως νευρωνικά δίκτυα και αλγόριθμοι ταξινόμησης. Η τεχνική που τελικά επιλέχθηκε για την πρόβλεψη της εξόδου κάποιου τελεστή, βασίζεται στους x κοντινότερους γείτονες του κειμένου που θέλουμε τα ταξινομήσουμε.

Β΄.5.1 Κοντινότεροι γείτονες

Για να προβλεφθεί η έξοδος ενός αλγορίθμου για ένα σύνολο κειμένων D_j με $D_j \subseteq D$, όπου το D περιέχει N κείμενα, λαμβάνουμε υπόψη τον *Similarity Matrix* ο οποίος είναι ένας πίνακας *[N*N]* που αποθηκεύει την ομοιότητα ($Similarity \in [0, 1]$) των N κειμένων ανά μεταξύ τους. Για κάθε κείμενο $j \in D_j$ βρίσκουμε στον *Similarity Matrix* τα x κείμενα που είναι πιο όμοια με αυτό, όπου αν $Similarity = 1$ τα δύο κείμενα είναι πανομοιότυπα και αν $Similarity = 0$ τα δύο κείμενα είναι τελείως ανόμοια. Βρίσκοντας τα πιο κοντινά κείμενα του κάθε j, εισάγουμε στον αλγόριθμο ένα τυχαίο σύνολο D_i, με $D_i \subseteq D$ και $D_i \cup D_j = D$, και αποθηκεύουμε το αποτέλεσμα του αλγορίθμου για κάθε $i \in D_i$. Έχοντας το αποτέλεσμα του αλγορίθμου για όλα τα $i \in D_i$, ελέγχουμε για κάθε $j \in D_j$ αν κάποια από τα x κοντινότερα κείμενα του ανήκουν στο D_i, δηλαδή αν έχει εφαρμοστεί ο τελεστής σε αυτά και άρα γνωρίζουμε το αποτέλεσμά του. Σε περίπτωση που κάποιο ή κάποια από τα x αυτά κείμενα ανήκουν στο D_i, έστω m τα κείμενα από τα x που ανήκουν στο D_i, ελέγχουμε την έξοδο του τελεστή για αυτό-α, το j ταξινομείται με το αποτέλεσμα που είχε ο τελεστής για τα περισσότερα κείμενα του m που είχαν την ίδια έξοδο, και το j προστίθεται στο D_i. Αν δεν υπάρχει κείμενο από

τα x που να ανήκει στο D_i, η διαδικασία συνεχίζεται για το επόμενο j. Για να γίνει πιο σαφές, έστω ότι λαμβάνουμε υπόψιν τα πέντε κοντινότερα κείμενα του j, δηλαδή $x = 5$, και σε αυτά τα πέντε υπάρχουν τρία που ανήκουν στο D_i και άρα γνωρίζουμε το αποτέλεσμα του αλγορίθμου για αυτά τα τρία. Αν τα δύο από τα τρία έχουν ίδιο αποτέλεσμα όταν εισαχθούν στον τελεστή, τότε και το j προβλέπεται ότι θα έχει την ίδια έξοδο. Η διαδικασία συνεχίζεται μέχρι να ισχύει ότι $D_i = D$, που σημαίνει ότι έχουμε προβλέψει την έξοδο του αλγορίθμου για όλα τα $j \in D_j$.

Δοκιμάστηκαν διάφοροι συνδυασμοί αριθμών για το x και το D_i, ώστε να υπάρχει μια ξεκάθαρη εικόνα για την αποτελεσματικότητα της μεθόδου. Πιο συγκεκριμένα, το x έλαβε τις τιμές $\{10, 15, 20, 1\% \times size(D), 3\% \times size(D)\}$ και το D_i $\{1, 3, 5, 10, 20\}\% \times size(D)$, όπου *size(D)* ο αριθμός των διαφορετικών κειμένων που εμπεριέχονται στο D. Για την δοκιμή της μεθόδου δοκιμάστηκαν όλοι οι πιθανοί συνδυασμοί του x με το D_i.

Β΄.5.2 Αλγόριθμος Ταξινόμησης

Μια άλλη τεχνική για την πρόβλεψη της εξόδου του αλγορίθμου που δοκιμάστηκε αλλά δεν ήταν αποτελεσματική σε ικανοποιητικό βαθμό, είναι ένα προεκπαιδευμένο μοντέλο πρόβλεψης με χρήση αλγορίθμου ταξινόμησης. Ο αλγόριθμος που χρησιμοποιήθηκε στο μοντέλο είναι το *Δέντρο Απόφασης*, και για την εκπαίδευση του χρησιμοποιήθηκαν αποστάσεις κειμένων μαζί με την έξοδο του τελεστή για το ένα από τα δύο. Πιο συγκεκριμένα, καθώς ο σκοπός της μεθόδου είναι να προβλεφθεί για ένα κείμενο $j \in D_j$ η έξοδος $F(j)$ ενός τελεστή, με δεδομένο την ομοιότητα του j με το κείμενο $i \in D_i$ και την έξοδο του τελεστή αν λάβει σαν είσοδο το i, το μοντέλο εκπαιδεύτηκε με τριάδες δεδομένων της μορφής $[Similarity_{i,j}, output_i, output_j]$ ώστε να προβλέπει το $output_j$.

Κατά την εφαρμογή της τεχνικής, βάσει του *Similarity Matrix*, για κάθε κείμενο $i \in D_i$ και $j \in D_j$ το μοντέλο παίρνει σαν είσοδο το $Similarity_{i,j}$ και το $output_i$, που στην προκειμένη περίπτωση είναι η γλώσσα που θεωρεί ο τελεστής πως είναι γραμμένο το i, ώστε να κάνει μία πρόβλεψη για το αν τα i, j είναι γραμμένα στην ίδια γλώσσα. Η πρόβλεψη είναι δυαδική(0, 1). Αν είναι θετική, το j ταξινομείται ότι είναι γραμμένο στην ίδια γλώσσα με το i και προστίθεται στο D_i , αλλιώς παραμένει στο D_j και συνεχίζεται η διαδικασία. Η πρόβλεψη σταματάει όταν ισχύει ότι $D_i = D$.

Β΄.6 Ανάλυση της Προτεινόμενης Μεθόδου

Έχοντας αναλύσει όλα τα επιμέρους στάδια της προτεινόμενης μεθόδου, πρέπει να γίνει μια συγκεντρωτική επεξήγηση της διαδικασίας. Η παρούσα εργασία, έχει ως στόχο την υλοποίηση μιας μεθόδου η οποία θα ταξινομεί αυτόματα δεδομένα κειμένου για τις ανάγκες κάποιας εργασίας, χωρίς να χρειάζεται το σύνολο(D) των δεδομένων να εισαχθούν στον εκάστοτε αλγόριθμο. Αυτό επιτυγχάνεται με την εφαρμογή του τελεστή σε ένα μικρό μέρος του συνόλου δεδομένων, και στην συνέχεια με βάση την έξοδο του προβλέπεται το αποτέλεσμα για τα υπόλοιπα δεδομένα.

Για τις ανάγκες της πρόβλεψης, πρέπει να υπάρξει κάποιο μέτρο συσχέτισης μεταξύ των δεδομένων. Σαν μέτρο συσχέτισης επιλέχθηκε η απόσταση ή αλλιώς ομοιότητα των κειμένων, που υποδηλώνει το πόσο απέχουν νοηματικά δύο κείμενα. Για να υπολογιστεί η απόσταση επιλέχθηκαν 3 αλγόριθμοι, όπου ο καθένας προσεγγίζει με διαφορετικό τρόπο την διαδικασία, και είναι η Ευκλείδεια απόσταση, το Cosine Distance και το Word Movers Distance. Έχοντας βρει τις αποστάσεις όλων των κειμένων του συνόλου δεδομένων μεταξύ τους, τις αποθηκεύουμε σε ένα πίνακα $SimilarityMatrix$ και στην συνέχεια εφαρμόζουμε τον επιλεγμένο αλγόριθμο σε ένα μικρό ποσοστό του συνόλου D, το D_i με $D_i \subseteq D$.

Ο τελεστής που επιλέχθηκε για να δοκιμαστεί η μέθοδος είναι ένας αλγόριθμος αναγνώρισης της γλώσσας ενός κειμένου. Το σύνολο των δεδομένων αποτελείται από μικρού μεγέθους κείμενα, γραμμένα σε 5 διαφορετικές λατινογενής γλώσσες(Αγγλικά, Γαλλικά, Γερμανικά, Ιταλικά, Ισπανικά). Έχοντας λοιπόν σαν δεδομένο τον $SimilarityMatrix$ και την γλώσσα που είναι γραμμένα τα κείμενα που ανήκουν στο D_i, προβλέπουμε την έξοδο του αλγορίθμου για τα υπόλοιπα κείμενα. Η πρόβλεψη γίνεται με βάση τους κοντινότερους γείτονες του κάθε κειμένου και την γλώσσα που είναι γραμμένοι αν ανήκουν στο D_j. Το κάθε κείμενο ταξινομείται στην γλώσσα που είναι η επικρατέστερη ανάμεσα στους κοντινότερους γείτονες του.

Η μέθοδος υλοποιήθηκε στην γλώσσα Python και το σύνολο δεδομένων αποτελούνταν από 10.000 διαφορετικά κείμενα. Για τις διάφορες δοκιμές, το υποσύνολο D_i πήρε τις τιμές $\{1, 3, 5, 10, 20\}\%$ του συνόλου των δεδομένων, και δοκιμάστηκε για τους $\{5, 10, 15, 20, 1\%, 3\%\}$ κοντινότερους γείτονες του κάθε κειμένου.

Για την δοκιμή της μεθόδου χρησιμοποιήθηκε VirtualMachine σε φορητό υπολογιστή αρχιτεκτονικής 64-bit, με παραχωρημένους 6 πυρήνες και 10GB μνήμης RAM. Το λειτουργικό σύστημα ήταν Ubuntu 21.0 και το λογισμικό για την συγγραφή και εκτέλεση του κώδικα το Visual Studio Code 1.59. Ο φορητός υπολογιστής διαθέτει συνολικά 8 πυρήνες, 16GB μνήμης RAM και λειτουργικό σύστημα Windows 10.

Β΄.7 Αποτελέσματα

Οι μέθοδος δοκιμάστηκε και για τους 3 αλγορίθμους απόστασης, με όλους τους πιθανούς συνδυασμούς του D_i με τους κοντινότερους γείτονες. Το σύνολο δεδομένων για το οποίο δοκιμάστηκε η μεθοδολογία αποτελούνταν από 10.000 κείμενα Τα συγκεντρωτικά αποτελέσματα των δοκιμών παρατίθενται παρακάτω:

Πίνακας Β΄.1: *Ακρίβεια της μεθόδου με την χρήση Euclidean Distance, σε ποσοστό%*

Euclidean Distance					
D_i	10-NN	15-NN	20-NN	1%-NN	3%-NN
1%	35.6	**36**	34.2	28	26
3%	**37**	34.5	31.8	31.6	28.7
5%	39.7	**40.5**	37.9	37.9	32.3
10%	43.8	41.7	**44.2**	39.4	35.2
20%	48.2	48.3	**52.7**	43.9	40.1

Πίνακας Β΄.2: *Ακρίβεια της μεθόδου με την χρήση Word Movers Distance, σε ποσοστό%*

Word Movers Distance					
D_i	10-NN	15-NN	20-NN	1%-NN	3%-NN
1%	**71**	58.8	61.2	65.7	55.2
3%	**85.5**	70.4	72	69.2	58.3
5%	75	72.5	**81**	70.8	65
10%	79.5	**81.8**	80	73.5	67.8
20%	**89**	86.5	85.5	88.5	87.7

Πίνακας Β΄.3: *Ακρίβεια της μεθόδου με την χρήση Cosine Similarity, σε ποσοστό%*

Cosine Similarity					
D_i	10-NN	15-NN	20-NN	1%-NN	3%-NN
1%	78.4	**84**	75.4	80	76
3%	86.6	81	86.5	**87**	78.9
5%	80.7	85.2	86	**87.3**	79.5
10%	88.4	**90**	89.3	88.3	82
20%	90.8	92.4	**93**	90.4	85.1

Για την κάθε μία διαφορετική περίπτωση η μέθοδος δοκιμάστηκε 5 φορές, και σαν τελική ακρίβεια θεωρήθηκε ο μέσος όρος των 5 δοκιμών. Στην παραπάνω ακρίβεια δεν συμπεριλαμβάνονται τα κείμενα που ανήκουν στο D_i. Αναφέρεται στο ποσοστό των κειμένων που δεν έχει εφαρμοστεί σε αυτά ο τελεστής, ή αλλιώς το ποσοστό κειμένων του D_j που προβλέφθηκε σωστά η γλώσσα που είναι γραμμένα. Επίσης, η ακρίβεια που αναφέρεται παραπάνω υπολογίστηκε συγκρίνοντας την έξοδο του τελεστή για το κείμενο και την πρόβλεψη της μεθόδου,

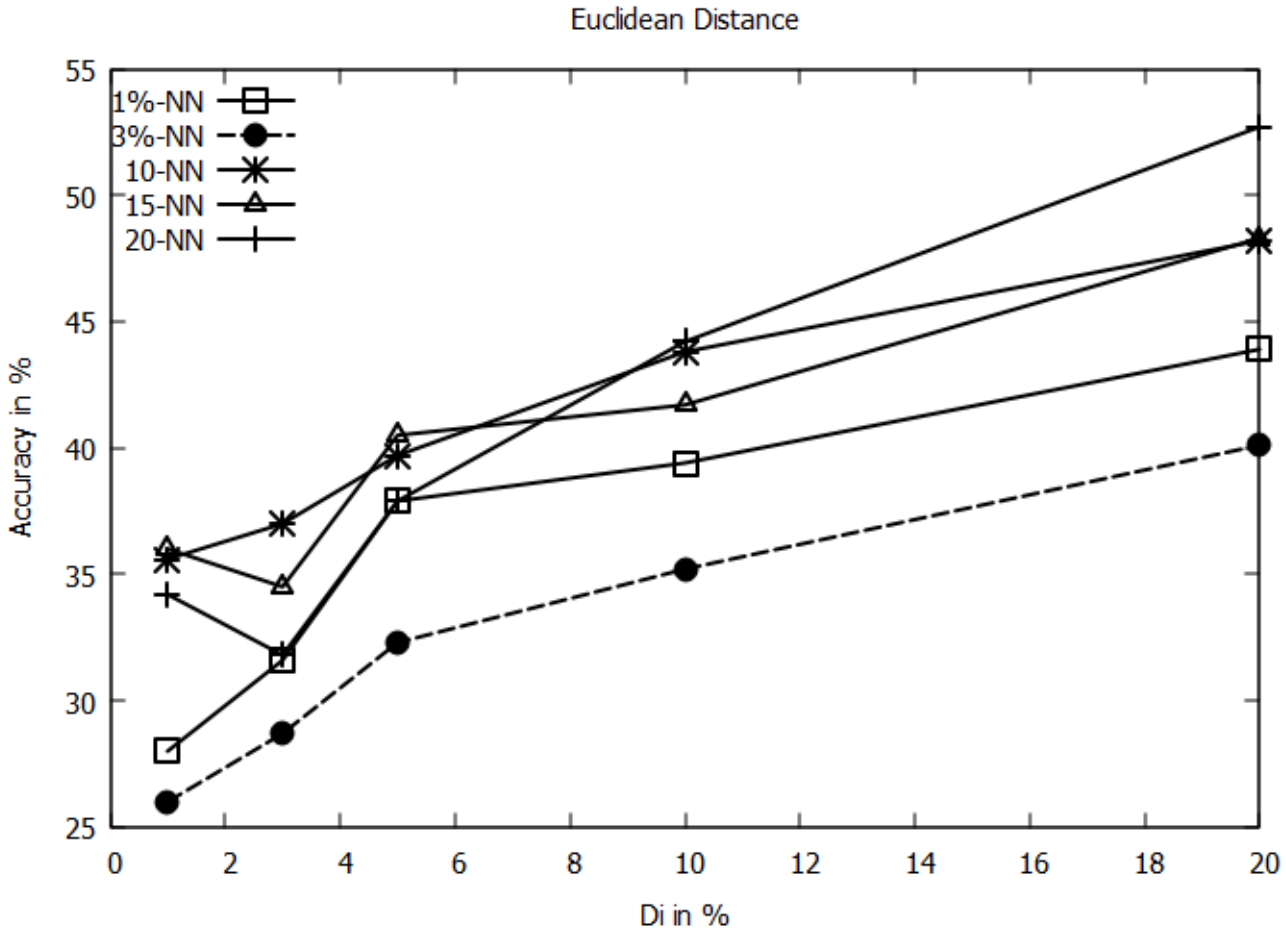
αλλά ο αλγόριθμος αναγνώρισης γλώσσας δεν είναι ακριβής στο 100%, έχει μια ακρίβεια της τάξης του 95%, οπότε η πραγματική ακρίβεια της μεθόδου είναι σε κάθε περίπτωση 0%-5% μεγαλύτερη.

Από τα αποτελέσματα και των τριών μέτρων απόστασης, φαίνεται πως όσο αυξάνεται ο αριθμός των κειμένων που εισάγονται αρχικά στον τελεστή, τόσο καλύτερη πρόβλεψη της εξόδου γίνεται και για τα υπόλοιπα κείμενα. Κάτι που είναι λογικό, αφού γνωρίζουμε για μεγαλύτερο αριθμό κειμένων την έξοδο του τελεστή.

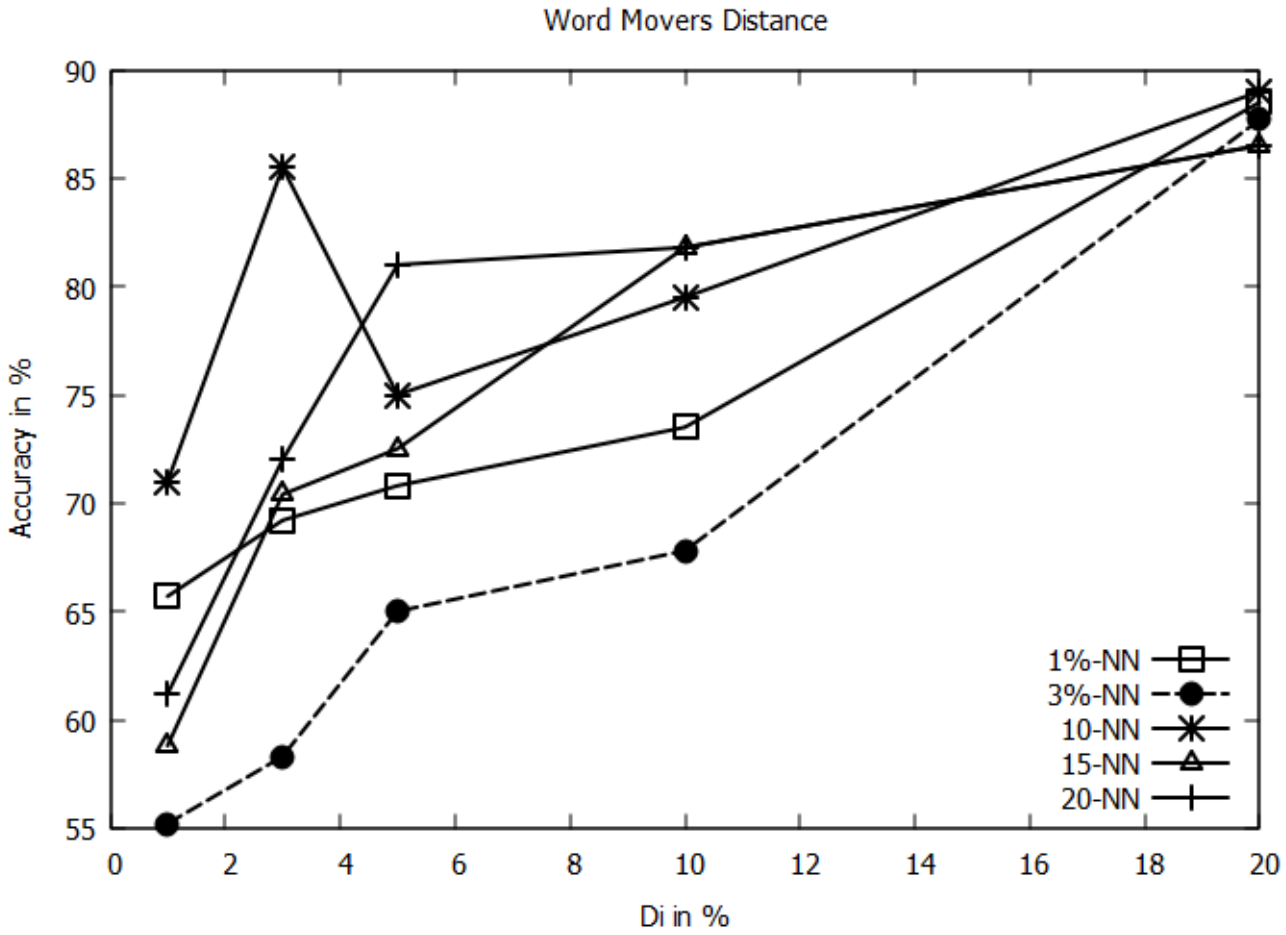
Σύμφωνα με τους παραπάνω πίνακες καθίσταται σαφές πως η πρόβλεψη της εξόδου ενός τελεστή για ένα μεγάλο αριθμό δεδομένων κειμένου, βάσει κάποιου μέτρου απόστασης, είναι έως ένα βαθμό εφικτή. Με μόλις το 3% του συνόλου δεδομένων να έχει εισαχθεί στον αλγόριθμο, προβλέφθηκε σωστά η έξοδος του για το 87% των κειμένων του υπολειπόμενου 97% του συνόλου, με την χρήση του Cosine Similarity. Με το μέγεθος του D_i να είναι το 20% του συνόλου δεδομένων η ακρίβεια είναι της τάξης του 93%.

Φαίνεται πως το μέτρο ομοιότητας που έχει την καλύτερη απόδοση, από την άποψη της ακρίβειας, είναι το Cosine Similarity, όπου για όλα τα διαφορετικά D_i έχει την μεγαλύτερη ακρίβεια στην πρόβλεψη των κειμένων του D_j. Αντίθετα, η Ευκλείδεια Απόσταση δεν έχει ικανοποιητική απόδοση, καθώς στην καλύτερη περίπτωση όπου το D_i είναι το 20% των κειμένων, η ακρίβεια των προβλέψεων είναι λίγο μεγαλύτερη από 50%, και στην περίπτωση όπου το D_i είναι το 1% έχει μέγιστη ακρίβεια μόλις 36%. Με την χρήση του αλγορίθμου Word Movers Distance, στην καλύτερη περίπτωση προβλέφθηκε σωστά το 89% των κειμένων ενώ στην χειρότερη η ακρίβεια είναι της τάξεως του 71%.

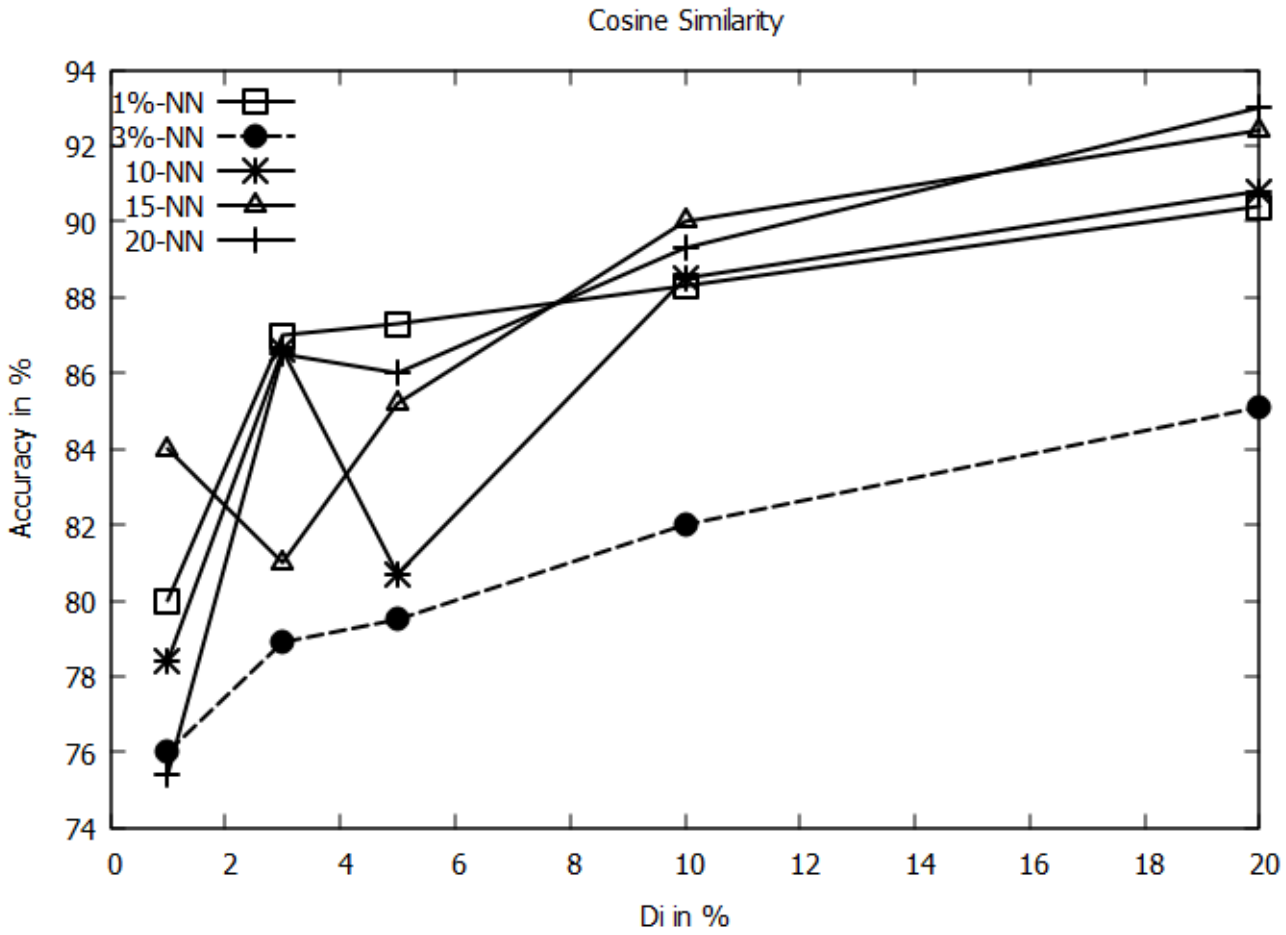
Β΄.7.1 Διαγράμματα



Σχήμα Β΄.2: *Ακρίβεια της μεθόδου με την χρήση Euclidean Distance, για όλα τα πιθανά D_i και NN*



Σχήμα Β΄.3: *Ακρίβεια της μεθόδου με την χρήση Word Movers Distance, για όλα τα πιθανά D_i και NN*



Σχήμα Β΄.4: *Ακρίβεια της μεθόδου με την χρήση Cosine Similarity, για όλα τα πιθανά D_i και NN*

Στα παραπάνω διαγράμματα, αναπαρίσταται η ακρίβεια της μεθόδου για κάθε έναν από τους αλγορίθμους απόστασης, σε συνδυασμό με το ποσοστό των κειμένων που εισάγονται στον τελεστή και για κάθε διαφορετικό αριθμό κοντινότερων γειτόνων όπου βασίζεται η πρόβλεψη.

Σύμφωνα με τις αναπαραστάσεις, δεν φαίνεται να υπάρχει κάποιος αριθμός κοντινότερων γειτόνων όπου να αποδίδει καλύτερα από τους υπόλοιπους. Αντίθετα, σε όλες τις περιπτώσεις υπάρχουν αντιστροφές στο ποιος συνδυασμός έχει την μεγαλύτερη ακρίβεια. Το μόνο που είναι ξεκάθαρο, είναι πως η περίπτωση όπου ο αριθμός των κοντινότερων γειτόνων ισούται με το 3% του συνόλου των δεδομένων έχει σχεδόν πάντα την χαμηλότερη ακρίβεια.

Με βάση όλα τα αποτελέσματα, δεν υπάρχει κάποιος ιδανικός συνδυασμός παραμέτρων όπου η μέθοδος αποδίδει καλύτερα. Αντιθέτως, όλες οι διαφορετικές παράμετροι αποδίδουν καλύτερα σε διαφορετικές περιπτώσεις. Τα μόνα ξεκάθαρα συμπεράσματα που εξάγονται, είναι πως το πιο αποδοτικό μέτρο απόστασης είναι το Cosine Similarity, και πως η χρήση του 3% των κοντινότερων γειτόνων για την πρόβλεψη δεν έχει ικανοποιητική ακρίβεια. Σίγουρα όμως θα μπορούσαμε να καταλήξουμε στο συμπέρασμα πως η μέθοδος στους καλύτερους συνδυασμούς παραμέτρων έχει σχετικά υψηλή ακρίβεια, όπου σίγουρα με διαφοροποιήσεις θα μπορούσε να αυξηθεί.

Β΄.8 Παράδειγμα Λειτουργίας

Για να αναλυθεί ένα ολοκληρωμένο παράδειγμα λειτουργίας πρέπει να υπάρξει κάποιο σύνολο δεδομένων.Θεωρούμε ένα σύνολο δεδομένων που αποτελείται από 20 τυχαία επιλεγμένα κείμενα από τα δεδομένα που συλλέχθηκαν για τις ανάγκες τις εργασίας, σε 5 διαφορετικές γλώσσες.

English

1. before the law was written down he was expected to memorise the laws and recite them from the law rock over the course of three summers
2. there are three islands in the lake
3. at the top of the food chain is the brown trout
4. this is due to the fact that when the weather is good it is usually best in this area

German

5. Die Verwendung dieses oder eines anderen Pseudonyms ist für Mitglieder der DGA streng reglementiert.
6. Ein Regisseur, der für einen von ihm gedrehten Film seinen Namen nicht hergeben möchte, hat nach Sichtung des fertigen Films drei Tage Zeit, anzuzeigen, dass er ein Pseudonym verwenden möchte.
7. Der Rat der DGA entscheidet binnen zwei Tagen über das Anliegen.
8. Erhebt die Produktionsfirma Einspruch, entscheidet ein Komitee aus Mitgliedern der DGA und der Vereinigung der Film- und Fernsehproduzenten, ob der Regisseur ein Pseudonym angeben darf.

French

9. les états-unis est généralement enneigée au printemps, et il est relaxant habituellement en juin.
10. paris est généralement froid pendant l' automne, et il gèle habituellement en septembre.
11. les états-unis est jamais pluvieux en février, mais il est beau en juillet.
12. Il y a un restaurant pas loin d'ici, tu pourrais y aller de temps en temps même si le personnel n'aime pas trop que vous sortiez

Spanish

13. Sus ideas radicales le valieron el sobrenombre de l anti tout el anti todo y de ahí ideó su seudónimo Lanti

14. defiende al barbero cuando es acosado por miembros de las fuerzas de seguridad de Hynkel

15. Ambos se enamoran y deben sufrir los atropellos de la dictadura

16. aunque esta era una de las películas predilectas que tenía Hitler en su cine particular

Italian

17. Questa definizione di Mitchell è rilevante poiché fornisce una definizione operativa dell'apprendimento automatico, invece che in termini cognitivi.

18. Fornendo questa definizione, Mitchell di fatto segue la proposta che Alan Turing fece nel suo articolo "Computing Machinery and Intelligence", sostituendo la domanda "Le macchine possono pensare?"

19. con la domanda "Le macchine possono fare quello che noi (in quanto entità pensanti) possiamo fare?".

20. L'obiettivo principe dell'apprendimento automatico è che una macchina sia in grado di generalizzare dalla propria esperienza[18], ossia che sia in grado di svolgere ragionamenti induttivi.

Έχοντας τα δεδομένα τα οποία θέλουμε να ταξινομήσουμε, πρώτο βήμα είναι να υπολογίσουμε την ομοιότητα όλων των δεδομένων μεταξύ τους. Για το συγκεκριμένο παράδειγμα σαν δείκτης ομοιότητας θα χρησιμοποιηθεί η απόσταση Cosine. Ο Similarity Matrix για τα συγκεκριμένα κείμενα με μέτρο την απόσταση Cosine είναι ο εξής:

Πίνακας Β΄.4: *Similarity Matrix*

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.0	0.708	0.513	0.742	1	1	1	1	1	1	1	1	1	1	1	1	1	0.971	1	1
2	–	0.0	0.724	0.789	1	1	1	1	1	1	1	1	1	1	1	1	0.915	1	0.910	0.874
3	–	–	0.0	0.594	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	–	–	–	0.0	1	1	1	1	1	1	1	1	1	1	1	1	0.958	1	0.956	0.938
5	–	–	–	–	0.0	0.908	0.777	0.749	1	1	1	1	1	1	1	1	1	1	1	1
6	–	–	–	–	–	0.0	0.904	0.705	1	1	1	1	1	1	1	1	1	1	1	1
7	–	–	–	–	–	–	0	0.566	1	1	1	1	1	1	1	1	1	1	1	1
8	–	–	–	–	–	–	–	0	1	1	1	1	1	1	1	1	1	1	1	1
9	–	–	–	–	–	–	–	–	0	0.509	0.428	0.906	1	1	1	0.937	1	1	1	1
10	–	–	–	–	–	–	–	–	–	0	0.659	0.888	1	1	1	0.925	1	1	1	1
11	–	–	–	–	–	–	–	–	–	–	0	0.863	1	1	1	0.878	1	1	1	1
12	–	–	–	–	–	–	–	–	–	–	–	0	0.884	0.874	0.939	0.850	1	0.964	0.954	1
13	–	–	–	–	–	–	–	–	–	–	–	–	0	0.738	0.873	0.845	1	0.962	0.952	1
14	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.792	0.774	1	1	1	1
15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.918	1	0.882	0.925	1
16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.942	1	1	0.956
17	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.750	0.894	0.767
18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.693	0.876
19	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0.842
20	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0

Ο παραπάνω πίνακας είναι συμμετρικός, γι΄ αυτό και υπολογίζεται κατά το ήμισυ. Οι γραμμές και οι στήλες του πίνακα αναφέρονται στις προτάσεις αντιστοίχως με την αρίθμηση που είχε γίνει παραπάνω. Επισημαίνεται πως όταν η απόσταση Cosine ισούται με 0 σημαίνει πως τα δύο κείμενα είναι πανομοιότυπα, και όταν ισούται με 1 σημαίνει πως είναι απόλυτα ανόμοια. Άρα, όσο πιο μικρή είναι η απόσταση τόσο πιο πολύ μοιάζουν τα δύο κείμενα.

Επόμενο βήμα είναι να εφαρμόσουμε τον εκάστοτε τελεστή σε ένα τυχαίο υποσύνολο του συνόλου δεδομένων. Ο αλγόριθμος που θα χρησιμοποιηθεί είναι ο ίδιος με αυτόν που χρησιμοποιήθηκε για την δοκιμή της μεθόδου, ο οποίος αναγνωρίζει την γλώσσα στην οποία είναι γραμμένο ένα κείμενο που του εισάγουμε. Το ποσοστό του συνόλου δεδομένων στο οποίο θα εφαρμοστεί ο αλγόριθμος, για λόγους ευκολίας, είναι το 25% ($D_i = 25\%$) ή αλλιώς 5 κείμενα. Καθώς η μεθοδολογία δεν θα λειτουργήσει ολοκληρωμένα για ένα τόσο μικρό σύνολο δεδομένων, αφού το πρόγραμμα εισέρχεται σε ατέρμων βρόγχο μην μπορώντας να ταξινομήσει όλα τα κείμενα, και σκοπός του παραδείγματος είναι η κατανόηση της μεθόδου, θα γίνει μια μικρή τροποποίηση και τα κείμενα στα οποία θα εφαρμοστεί ο αλγόριθμος δεν θα είναι εντελώς τυχαία αλλά ένα από κάθε διαφορετική γλώσσα. Τα κείμενα τα οποία επιλέχθηκαν για να εφαρμοστεί σε αυτά ο αλγόριθμος είναι τα {1,7,11,13,19} και τα αποτελέσματα του αλγορίθμου για τα παραπάνω κείμενα είναι {en,de,fr,es,it} αντίστοιχα.Όπου:

- en⟶ Αγγλικά
- de⟶ Γερμανικά
- fr⟶ Γαλλικά
- es⟶ Ισπανικά
- it⟶ Ιταλικά

Αφού έχουμε ταξινομήσει ένα υποσύνολο του συνόλου δεδομένων, συνεχίζουμε ταξινομώντας τα υπόλοιπα κείμενα με την μέθοδο των κοντινότερων γειτόνων.Για το συγκεκριμένο παράδειγμα θα λάβουμε υπόψη τον 1 κοντινότερο κείμενο κάθε εγγραφής, ή αλλιώς 1-NN. Ξεκινώντας από το κείμενο νούμερο 2, αφού το 1 έχει ήδη ταξινομηθεί, βάσει του Similarity Matrix φαίνεται πως το κείμενο με την μικρότερη απόσταση από αυτό είναι το νούμερο 1. Το νούμερο 1 έχει ταξινομηθεί με την γλώσσα 'Αγγλικά΄, οπότε και το 2 ταξινομείται με την γλώσσα 'Αγγλικά΄. Για το νούμερο 3 το κοντινότερο κείμενο είναι πάλι το 1, οπότε και το νούμερο 3 ταξινομείται σαν 'Αγγλικα΄. Για το κείμενο 4 το κοντινότερο κείμενο είναι το νούμερο 3, το οποίο ταξινομήθηκε στο προηγούμενο βήμα σαν 'Αγγλικά΄ οπότε και το 4 ταξινομείται σαν 'Αγγλικά΄. Για το κείμενο 5 το κοντινότερο είναι το 8 το οποίο δεν έχει ταξινομηθεί, οπότε η ταξινόμηση συνεχίζεται με το επόμενο. Η ίδια διαδικασία επαναλαμβάνεται μέχρι να ταξινομηθούν όλα τα κείμενα. Στον πίνακα Β΄ 5 αναγράφονται αναλυτικά όλα τα βήματα για την ταξινόμηση των κειμένων.

Πίνακας Β΄.5: *Βήματα Ταξινόμησης*

Βήμα	Κείμενο	Κοντινότερος Γείτονας	Γλώσσα Κοντινότερου Γείτονα	Ταξινομημένα Κείμενα {1-20}
0	{1,7,11,13,19}	-	-	{en,-,-,-,-,de,-,-,-,-,fr,-,es,-,-,-,-,-,it,-}
1	2	1	en	{en,en,-,-,-,de,-,-,-,-,fr,-,es,-,-,-,-,-,it,-}
2	3	1	en	{en,en,en,-,-,-,de,-,-,-,fr,-,es,-,-,-,-,-,it,-}
3	4	3	en	{en,en,en,en,-,-,de,-,-,-,fr,-,es,-,-,-,-,-,it,-}
4	5	8	-	{en,en,en,en,-,-,de,-,-,-,fr,-,es,-,-,-,-,-,it,-}
5	6	8	-	{en,en,en,en,-,-,de,-,-,-,fr,-,es,-,-,-,-,-,it,-}
6	8	7	de	{en,en,en,en,-,-,de,de,-,-,fr,-,es,-,-,-,-,-,it,-}
7	9	11	fr	{en,en,en,en,-,-,de,de,fr,-,fr,-,es,-,-,-,-,-,it,-}
8	10	9	fr	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,-,-,-,-,-,it,-}
9	12	16	-	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,-,-,-,-,-,it,-}
10	14	13	es	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,-,-,-,-,it,-}
11	15	14	es	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,es,-,-,-,it,-}
12	16	14	es	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,es,es,-,-,it,-}
13	17	18	-	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,es,es,-,-,it,-}
14	18	19	it	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,es,es,-,it,it,-}
15	20	17	-	{en,en,en,en,-,-,de,de,fr,fr,fr,-,es,es,es,es,-,it,it,-}
16	5	8	de	{en,en,en,en,de,-,de,de,fr,fr,fr,-,es,es,es,es,-,it,it,-}
17	6	8	de	{en,en,en,en,de,de,de,de,fr,fr,fr,-,es,es,es,es,-,it,it,-}
18	12	16	es	{en,en,en,en,de,de,de,de,fr,fr,fr,es,es,es,es,es,-,it,it,-}
19	17	18	it	{en,en,en,en,de,de,de,de,fr,fr,fr,es,es,es,es,es,it,it,it,-}
20	20	17	it	{en,en,en,en,de,de,de,de,fr,fr,fr,es,es,es,es,es,it,it,it,it}

Στο παραπάνω παράδειγμα η μεθοδολογία ταξινόμησε σωστά όλα τα κείμενα του συνόλου δεδομένων. Με τον ίδιο τρόπο γίνεται και η ταξινόμηση μεγαλύτερης κλίμακας συνόλων, με την

μόνη διαφορά ότι λαμβάνεται υπόψη μεγαλύτερος αριθμος κοντινών κειμένων και το εκάστοτε κείμενο ταξινομείται σύμφωνα με την πλειοψηφία.

Κεφάλαιο Γ΄

Συμπεράσματα

Με την επιτυχή υλοποίηση της παρούσας εργασίας, εξήχθησαν πολλά συμπεράσματα που αφορούν την διαχείριση και επεξεργασία δεδομένων φυσικής γλώσσας, τους αλγορίθμους ομοιότητας, τους τελεστές που εφαρμόζονται σε δεδομένα φυσικής γλώσσας, καθώς και τις ιδιαιτερότητες των ανεπεξέργαστων δεδομένων που συλλέγονται από το διαδίκτυο. Ο συνδυασμός όλων αυτών των πρακτικών επέφερε πολλές προκλήσεις, αλλά ταυτόχρονα και πολλά χρήσιμα διδάγματα.

Το πιο βασικό κομμάτι μιας εργασίας η οποία βασίζεται στην ανάλυση δεδομένων, είναι καταρχάς, η συλλογή των κατάλληλων δεδομένων και μετέπειτα η σωστή επεξεργασία τους ώστε να μπορέσουν να εξαχθούν συμπεράσματα από αυτά. Είναι σχεδόν απίθανο να βρεθούν δεδομένα τα οποία είναι ανεπεξέργαστα αλλά ταυτόχρονα και με την δομή που χρειαζόμαστε για τις ανάγκες της εργασίας. Πάντα, στα μεγάλα σύνολα δεδομένων που συλλέγονται από πηγές θα υπάρχει θόρυβος, αχρείαστη πληροφορία αλλά και πολλά λάθη, τα οποία δυσχεραίνουν την διαδικασία της ανάλυσης. Είναι λοιπόν πολύ σημαντικό να γνωρίζουμε ακριβώς τον τελικό στόχο της εκάστοτε εργασίας, ώστε να μπορέσουμε να συγκεντρώσουμε όσο το δυνατόν καλύτερα δεδομένα και να τα επεξεργαστούμε κατάλληλα. Οι ίδιες προκλήσεις συναντώνται και κατά την ανάλυση δεδομένων φυσικής γλώσσας, με μόνη διαφορά ότι σε αυτή την περίπτωση υπάρχουν και άλλα εμπόδια τα οποία πρέπει να ξεπεραστούν, και πηγάζουν κυρίως από τις ιδιαιτερότητες που έχουν ο προφορικός και γραπτός λόγος. Η φυσική γλώσσα, έχει κάποια χαρακτηριστικά τα οποία καθιστούν την ανάλυση της ιδιαίτερα δύσκολη διαδικασία. Για παράδειγμα, όταν μιλάμε ή γράφουμε μπορούμε με πολλές διαφορετικές προτάσεις να αποδώσουμε το ίδιο νόημα, πολλές φορές πράγματα αυτονόητα παραλείπονται, υπάρχουν πολλές συνώνυμες λέξεις και γίνεται συχνά χρήση λαϊκών εκφράσεων. Ειδικότερα όταν καλούμαστε να αναλύσουμε δεδομένα που προέρχονται από κείμενα καθημερινών ανθρώπων όλα τα παραπάνω συναντώνται σε μέγιστο βαθμό. Όλα αυτά συντελούν στο να είναι η εξαγωγή γνώσης από δεδομένα κειμένου μια εξαιρετικά πολύπλοκη διαδικασία, ταυτόχρονα όμως η εκμετάλλευση δεδομένων σε φυσική γλώσσα είναι πλέον απαραίτητη καθώς ο όγκος τους είναι τεράστιος και η πληροφορίες που μπορούν να παρέχουν κομβικής σημασίας.

Τα τελευταία χρόνια, η πλειονότητα των δεδομένων φυσικής γλώσσας που συλλέγεται και επεξεργάζεται προέρχεται από το διαδίκτυο, και συνήθως σκοπός της ανάλυσης είναι η

εξαγωγή συμπερασμάτων για τους χρήστες του. Αναμφισβήτητα τα μέσα κοινωνικής δικτύωσης αποτελούν μια τεράστια πηγή παροχής δεδομένων, κυρίως κειμένων, η οποία συνεχώς επεκτείνεται και ανανεώνεται με ταχύτατους ρυθμούς. Μέσα από δημοσιεύσεις, μηνύματα και σχόλια των χρηστών μπορούμε να δούμε την άποψη της κοινής γνώμης πάνω σε κάποιο θέμα και να κρίνουμε το αν υπάρχει θετική ή αρνητική αξιολόγηση για αυτό, με το θέμα να είναι οτιδήποτε από ένα προϊόν ή μία υπηρεσία, μέχρι πρόσωπα και κοινωνικές καταστάσεις. Μια άλλη διαδικασία που βασίζεται στην ανάλυση δεδομένων φυσικής γλώσσας είναι αυτή του διαχωρισμού κειμένων είτε με βάση την θεματολογία τους, είτε με βάση κάποιο χαρακτηριστικό τους (π.χ τον συγγραφέα, την αξιοπιστία, το ύφος κ. α). Τέτοιες εφαρμογές χρησιμοποιούνται για τον εντοπισμό ψευδών ειδήσεων, ανεπιθύμητων μηνυμάτων και παραπλανητικών δημοσιεύσεων. Τέλος, μια ακόμα επίκαιρη εφαρμογή που επεξεργάζεται δεδομένα φυσικής γλώσσας είναι οι αυτόματοι ομιλητές και οι ψηφιακοί βοηθοί. Υπάρχουν πολλές ακόμα εφαρμογές που βασίζονται στην ανάλυση δεδομένων κειμένου, αλλά οι προαναφερθείσες είναι κάποιες από τις κυριότερες.

Η ομοιότητα μεταξύ δύο κειμένων μπορεί να υπολογιστεί με διάφορους τρόπους και σύμφωνα με διαφορετικά στοιχεία τους, πάντα όμως το πιο βασικό είναι οι κοινές ή οι συνώνυμες λέξεις τους. Είναι προφανές πως οι κοινές λέξεις είναι ένα βασικό στοιχείο δύο προτάσεων ώστε να χαρακτηριστούν όμοιες, αλλά το πρόβλημα σε αυτή την προσέγγιση, βάσει τους στόχους της παρούσας εργασίας, είναι ότι δεν αναγνωρίζεται η συναισθηματική ομοιότητα και η κοινή θεματολογία δύο κειμένων. Αυτό συμβαίνει γιατί πρώτων, τα συναισθήματα που εκφράζει ο συγγραφέας μέσα από το κείμενο δεν είναι άρρηκτα συνδεδεμένα με το λεξιλόγιο που χρησιμοποιεί, και γιατί κείμενα με την ίδια θεματολογία δεν χρησιμοποιούν απαραίτητα παρεμφερείς λέξεις. Για παράδειγμα δύο ιατρικά κείμενα που αναφέρονται σε διαφορετικές ασθένειες το πιο πιθανό είναι πως χρησιμοποιούν διαφορετικές ορολογίες και άρα θα έχουν μικρή ομοιότητα, παρόλο που και τα δύο υπάγονται στον κλάδο της ιατρικής. Γιάυτό και είναι περιορισμένοι οι αλγόριθμοι των οποίων η έξοδος μπορεί να προβλεφθεί σύμφωνα με κάποιο δείκτη ομοιότητας.

Όσον αφορά συνολικά την παρούσα υλοποίηση, βγαίνει το συμπέρασμα ότι η πρόβλεψη της εξόδου ενός αλγορίθμου που εφαρμόζεται σε δεδομένα κειμένου με κριτήριο την απόσταση μεταξύ των δεδομένων εισόδου είναι εφικτή, αλλά σίγουρα το εύρος των πιθανών τελεστών είναι περιορισμένο. Λόγω των ιδιαιτεροτήτων που χαρακτηρίζουν την φυσική γλώσσα, η προτεινόμενη μεθοδολογία δεν θα ήταν επιτυχής για οποιονδήποτε αλγόριθμο αλλά για όσους έχουν, τις περισσότερες φορές, ίδιο αποτέλεσμα για 2 κείμενα με αρκετές ίδιες ή συνώνυμες λέξεις. Το πιο βασικό κομμάτι για την επιτυχία της μεθόδου είναι το μέτρο απόστασης που θα επιλεγεί. Πρέπει να γίνουν ξεκάθαρα τα κριτήρια που χαρακτηρίζουν δύο κείμενα όμοια, καθώς αυτά τα κριτήρια είναι αλληλοσχετιζόμενα με τον τελεστή για τον οποίον θα χρησιμοποιηθεί η προτεινόμενη μεθοδολογία. Στην παρούσα εργασία προτάθηκε ένας αλγόριθμος για τον οποίο υπήρξε υψηλή ακρίβεια, αλλά σίγουρα θα μπορούσε να έχει αντίστοιχα ή και καλύτερα αποτελέσματα και για άλλους. Ένα άλλο σημείο που θα πρέπει να ληφθεί υπόψη κατά την επιλογή του κατάλληλου τελεστή είναι η χρονική πολυπλοκότητα του, καθώς στόχος της παρούσας εργασίας είναι να προτείνει μια μέθοδο ώστε να μειωθεί ο χρόνος που απαιτείται για να ¨τρέξει' ένας αλγόριθμος, οπότε κάτι τέτοιο δεν θα είχε νόημα για αλγορίθμους που έχουν χαμηλή χρονική πολυπλοκότητα.

Βιβλιογραφία

[1] Jianqing Fan, Fang Han, Han Liu, *Challenges of Big Data analysis* , National Science Review 1: 293–314, 2014

[2] Hossein Hassani , Xu Huang and Emmanuel Silva, *Digitalisation and Big Data Mining in Banking*, Big Data and Cognitive Computing, 2018

[3] Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa, *Data Mining in Healthcare for Heart Diseases* ,International Journal of Innovation and Applied Studies,2015

[4] P. Ahmad, S. Qamar, S. Q. A. Risvi, *Techniques of Data Mining In Healthcare: A Review*,International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2015

[5] NICHOLAS M. BALL, ROBERT J. BRUNNER *DATA MINING AND MACHINE LEARNING IN ASTRONOMY*,International Journal of Modern Physics D, 2010

[6] S. Filippov, *Mapping Text and Data Mining in Academic and Research Communities in Europe* , Technical Report,2014

[7] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

[8] https://www.statista.com/

[9] https://twitter.com/

[10] D. Bahdanau, K. Cho, Y. Bengio, *Neural machine translation by jointly learning to align and translate*, International Conference on Learning Representations, 2015

[11] M Sasaki, H Shinnou, *Spam Detection Using Text Clustering*, 2005 International Conference on Cyberworlds (CW'05). IEEE, 2005

[12] Xu, Qian, et al., *SMS Spam Detection Using Non-Content Features*, IEEE Intelligent Systems 27.6 (2012): 44-51

[13] M. Dahiya, *A tool of conversation: Chatbot*, International Journal of Computer Sciences and Engineering 5.5 (2017): 158-161.

[14] P. Kumar, M. Sharma, S. Rawat and T. Choudhury,, *Designing and Developing a Chatbot Using Machine Learning*, 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), 2018, pp. 87-91

[15] Medhat, W., Hassan, A., Korashy, H. , *Sentiment analysis algorithms and applications: A survey*, Ain Shams engineering journal, 5(4), 1093-1113.

[16] Nenkova, A., McKeown, K., *Automatic summarization.*, Now Publishers Inc, 2011

[17] D.Tsoumakos, I.Giannakopoulos *Content-based Analytics: Moving Beyond Data Size* ,2020 IEEE Sixth International Conference on Big Data Computing Service and Applications

[18] https://pypi.org/project/stop-words/

[19] P.Willett, *The Porter stemming algorithm: then and now* , Program: electronic library and information systems, 40 (3). pp. 219-223, 2006

[20] https://wordnet.princeton.edu/

[21] *Understanding TF-ID: A Simple Introduction* https://monkeylearn.com/blog/what-is-tf-idf/

[22] J. Pennington, R. Socher, C.D. Manning , *GloVe: Global Vectors for Word Representation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages(1532–1543)

[23] Stalker, George H. , *Some Notions of 'Similarity' among Lines of Text.* Computers and the Humanities, vol. 11, no. 4, Springer, 1977, pp. 199–209, http://www.jstor.org/stable/30199897.

[24] D. Gunawan, C. A. Sembiring, M. A. Budiman , *The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents* et al 2018 J. Phys.: Conf. Ser. 978 012120

[25] A. ISLAM and D. INKPEN , *Semantic text similarity using corpus-based word similarity and string similarity* . ACM Trans. Knowl. Discov. Data. 2, 2, Article 10 (July 2008),

[26] N. Pradhan,M. Gyanchandani,R. Wadhvani , *A Review on Text Similarity Technique used in IR and its Application* International Journal of Computer Applications (0975 – 8887) Volume 120 – No.9, June 2015

[27] M. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, *From Word Embeddings To Document Distances* Proceedings of the 32nd International Conference on Machine Learning, 2015

[28] T. Mikolov, K. Chen, G. Corrado, J. Dean , *Efficient Estimation of Word Representations in Vector Space* arXiv:1301.3781v3 [cs.CL] 7 Sep 2013

[29] https://radimrehurek.com/gensim/

[30] https://pypi.org/project/langdetect/

[31] https://www.kaggle.com/henryhaefliger/english-sentences-with-noise?select=clean.txt

[32] https://www.kaggle.com/bminixhofer/8m-german-sentences-from-wikipedia

[33] https://zenodo.org/record/4319957#.YXW1UhpBxPZ

[34] https://www.kaggle.com/alonyoeli/twitter-italian-dialect-data

[35] https://www.kaggle.com/faouzimohamed/englishfrench-fornmt?select=english-french.csv