Πανεπιστήμιο Πελοποννήσου
Τμήμα Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΔΕΔΟΜΕΝΑ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Παναγάκη Ιωάννα | Α.Μ. 2622 | Ημερομηνία
Φωτάκια Διονυσία | Α.Μ. 2620 | Ημερομηνία

Επιβλέπον καθηγητής: Μιχαήλ Παρασκευάς

Πάτρα-Φεβρουάριος 2020

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή
Πάτρα, 24/02/2020

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Μιχαήλ Παρασκευάς
2. Σωτήριος Χριστοδούλου
3. Ιωάννης Τζήμας

Περιεχόμενα

Εισαγωγή

Στην παρούσα πτυχιακή εργασία ασχοληθήκαμε με την ανάλυση συναισθήματος σε δεδομένα κοινωνικών δικτύων με μεθόδους μηχανικής μάθησης.

Την ανάλυση συναισθήματος τη χωρίσαμε σε τρεις κατηγορίες: θετικό, αρνητικό και ουδέτερο συναίσθημα.

Για την διεξαγωγή της εργασίας αυτής χρησιμοποιήσαμε 1340 περίπου σχόλια χρηστών από αναρτήσεις που είχαν γίνει στα κοινωνικά δίκτυα, ένα από τα πιο δημοφιλή λογισμικά μηχανικής μάθησης το **Weka** (Waikato Environment for Knowledge Analysis) καθώς επίσης και μερικές εντολές της γλώσσας προγραμματισμού **Python.**

Την ανάλυση συναισθήματος σαν πρόβλημα μπορούσαμε να το προσεγγίσουμε με διαφορετικούς τρόπους, ωστόσο εμείς στη συγκεκριμένη εργασία προσεγγίσαμε το πρόβλημα αυτό με τη μηχανική μάθηση (machine learning).

Η μηχανική μάθηση δίνει στους υπολογιστές (μηχανές) την δυνατότητα να μαθαίνουν με έναν τρόπο που μοιάζει με την λειτουργία του ανθρώπινου εγκεφάλου.

Abstract

In the present thesis, we have dealt with the analysis of emotionality in social networking data using machine learning methods.

The analysis of emotionality is divided into three categories: positive, negative and neutral.

About 1340 user feedback from social media postings, one of the most popular Weka (Waikato Environment for Knowledge Analysis) machine learning software, and some Python language programming commands were used to carry out this work.

We could approach emotion analysis as a problem in different ways, but in this work we approached this problem with machine learning.

Machine learning enables computers (machines) to learn in a way that resembles the function of the human brain.

Κεφάλαιο 1: Ανάλυση συναισθήματος

1.1 Εισαγωγή στην ανάλυση συναισθήματος

Η ανάλυση συναισθήματος είναι ένα συνεχώς αναπτυσσόμενο πεδίο της μηχανικής μάθησης, αποτελεί τομέα της επεξεργασίας φυσικής γλώσσας (NLP) και στόχο έχει την κατηγοριοποίηση κειμένων με βάση την πολικότητά τους. Σε αυτή την εργασία θα γίνει κατηγοριοποίηση φράσεων και όχι ολόκληρων κειμένων. Θα μπορούσαμε να κατηγοριοποιήσουμε τη πολικότητα των φράσεων αυτών σε μία κλίμακα πολλών κατηγοριών, αλλά επιλέξαμε να τη κατηγοριοποιήσουμε με το χαρακτηρισμό θετικό, αρνητικό και ουδέτερο.

1. 2 Κατηγορίες προσέγγισης κειμένου

Για να είναι δυνατή η εξαγωγή ανάλυσης συναισθήματος, υπάρχουν αλγόριθμοι που μπορούν να κατηγοριοποιηθούν με βάση τον τρόπο που προσεγγίζουμε το κείμενο ή τις φράσεις και διάφορες τεχνικές επεξεργασίας της φυσικής γλώσσας οι οποίες αναλύονται παρακάτω.

1.2.1 Ταξινόμηση σε επίπεδο κειμένου

Αυτή η προσέγγιση θεωρεί ότι κάθε κείμενο περιέχει τις απόψεις ενός μόνο ατόμου γύρω από ένα συγκεκριμένο θέμα και έχει ως στόχο να το χαρακτηρίσει το συναίσθημα που εκφράζεται μέσα από το κείμενο ως θετικό ή αρνητικό. Οι περισσότερες τεχνικές ανάλυσης συναισθήματος κειμένου είναι εποπτευόμενης μάθησης αλλά υπάρχουν και τεχνικές μη εποπτευόμενης μάθησης, έννοιες που θα αναλύσουμε στο κεφάλαιο 2.2 .

1.2.1.1 Κατηγοριοποίηση κειμένων

Στην ανάλυση συναισθήματος υπάρχουν δύο κύριοι οδοί έρευνας για την κατηγοριοποίηση κειμένων:

1^{ος} Η λεξιλογική προσέγγιση, που εστιάζει στην οικοδόμηση επιτυχημένων λεξικών και μπορεί να αναφέρεται:

α) στη γενικότερη συναισθηματική κατάσταση του συγγραφέα κατά τη συγγραφή του κειμένου.

β) στο συναίσθημα που μεταδίδεται σκόπιμα από τον συγγραφέα στον αναγνώστη μέσω του κειμένου και

γ) στην στάση – άποψη – εκτίμηση του συγγραφέα σχετικά με κάποιο θέμα.

2ος Η προσέγγιση μηχανικής μάθησης, η οποία εστιάζει στα διανύσματα χαρακτηριστικών γνωρισμάτων και με αυτόν τον τρόπο μπορεί να αναζητείται το συναίσθημα ή η πολικότητά του σε:

- ένα ολόκληρο κείμενο
- μία πρόταση ή ακόμα και σε
- μεμονωμένες φράσεις.

1.2.2 Ταξινόμηση σε επίπεδο πρότασης

Η ταξινόμηση αυτή εστιάζει στη πρόταση και τον ακριβή προσδιορισμό της θετικής, αρνητικής ή ουδέτερης στάσης που εκφράζει. Επιπλέον συχνά συνδέεται με την ταξινόμηση υποκειμενικότητας, που διαχωρίζει τις προτάσεις ανάλογα με το αν περιέχουν γεγονότα της πραγματικότητας ή προσωπικές απόψεις. Τα κριτήρια διαχωρισμού των προτάσεων αποτελούν τις δύο κλάσεις της ταξινόμησης υποκειμενικότητας οι οποίες ονομάζονται αντικειμενική για το πρώτο κριτήριο και υποκειμενική για το δεύτερο κριτήριο διαχωρισμού.

1.2.3 Ταξινόμηση σε επίπεδο λέξης

Το επίπεδο αυτό χρησιμοποιείτε για ταξινόμηση επιπέδου πρότασης ή κειμένου και βασίζεται στο ότι οι πιο σημαντικοί δείκτες συναισθημάτων είναι οι λέξεις. Μία λίστα από τέτοιες λέξεις ονομάζεται λεξικό συναισθημάτων. Για τη δημιουργία λεξικών συναισθημάτων χρησιμοποιούνται πληροφορίες που προκύπτουν από την επεξεργασία κειμένων ή από γλωσσολογικούς πόρους όπως τα λεξικά.

1.2.4 Ταξινόμηση σε επίπεδο οντότητας και χαρακτηριστικών

Η εν λόγω ταξινόμηση βασίζεται κυρίως στην ιδέα ότι μία υποκειμενική κρίση αποτελείται από ένα συναίσθημα και ένα στόχο στον οποίο απευθύνεται και ο οποίος στα περισσότερα συστήματα κειμενικής ανάλυσης αναπαρίσταται μέσω οντοτήτων.

1.3 Προβλήματα της ανάλυσης συναισθήματος

Υπάρχουν προβλήματα που δεν μας επιτρέπουν να κάνουμε σωστά την ανάλυση συναισθήματος . Αυτά τα προβλήματα μπορεί να είναι το μήκος του κειμένου ή το πολυγλωσσικό περιεχόμενο ακόμα και ότι το συναίσθημα μπορεί πολλές φορές να εκφραστεί με πιο έμμεσο τρόπο. Επιπλέον ας μην ξεχνάμε πως ο προσδιορισμός του εκφραστή της άποψης που διατυπώνεται στο κείμενο μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκύψει διαφορετικό συναίσθημα από τη διαφορετική σειρά των λέξεων.

1.3.1 Προβλήματα της ανάλυσης συναισθήματος με βάση τη τεχνική που εφαρμόζουμε

Ανάλογα με την τεχνική που εφαρμόζουμε προκύπτουν και τα αντίστοιχα προβλήματα που διακρίνονται παρακάτω.

Προσεγγίσεις με λεξικά:

- Το πλήθος των λέξεων είναι πεπερασμένο, σε αντίθεση για ανάλυση σε ένα δυναμικό μέσο.
- Λέξεις (ίδιες) που μπορούν να έχουν διαφορετικό συναίσθημα σε διαφορετικά περιεχόμενα.

Προσεγγίσεις με εποπτευόμενη μηχανική μάθηση:

- Το κόστος.
- Η αδυναμία να εκπαιδευτούν οι ταξινομητές σε περισσότερα από ένα πεδία.
- Ο πολύ μεγάλος χρόνος και κόπος που απαιτείτε για την κατασκευή του συνόλου εκπαίδευσης του ταξινομητή.
- Η δυσκολία στην εύρεση αντιπροσωπευτικού δείγματος για την κατάλληλη εκπαίδευση του ταξινομητή.

1.4 Χαρακτηριστικά της ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος εφαρμόζεται για κατανόηση του νοήματος ενός κειμένου. Πολλοί αναφέρονται σε αυτό ως ένα σύγχρονο εργαλείο που χρησιμοποιείτε για την επεξεργασία και τη λήψη αποφάσεων μέσω των θετικών ή αρνητικών απόψεων των ανθρώπων. Η κύρια μεθοδολογία που χαρακτηρίζει την ανάλυση συναισθήματος είναι η Συναισθηματική Κατηγοριοποίηση.

1.5 Τρόποι προσέγγισης του προβλήματος της ανάλυσης συναισθήματος

Ο Τρόπος που προσεγγίσαμε την ανάλυση συναισθήματος σαν πρόβλημα είναι η μηχανική μάθηση ωστόσο υπάρχουν και άλλοι τρόποι προσέγγισης όπως:

Α) **Η προσέγγιση βασισμένη σε λεξικό** (lexicon-based method), η οποία σε αντίθεση με τις μεθόδους βασισμένες σε μηχανική μάθηση δεν απαιτεί την εκπαίδευση ενός ταξινομητή. Αντιθέτως χρησιμοποιεί λεξικά συναισθήματος για να αποδώσει το συναίσθημα των συναισθηματικά φορτισμένων λέξεων στο κείμενο. Η απόδοση μιας μεθόδου Ανάλυσης Συναισθήματος βασισμένη σε λεξικό συνήθως καθορίζεται από τον τύπο του λεξικού συναισθήματος και από τον αλγόριθμο που χρησιμοποιείται για τον εντοπισμό των συναισθηματικά φορτισμένων λέξεων του κειμένου και τον υπολογισμό του συνολικού συναισθήματος.

Β) **Η υβριδική προσέγγιση ,** η οποία στην ουσία προσπαθεί να εκμεταλλευτεί και να ενώσει τα πλεονεκτήματα των δύο μεθόδων και να αποφύγει τα μειονεκτήματα. Γενικότερα στις υβριδικές προσεγγίσεις η μία από τις δύο βασικές προσεγγίσεις χρησιμοποιείται για να τονώσει την απόδοση της άλλης προσέγγισης.

1.5.1 Σύγκριση τρόπων προσέγγισης του προβλήματος της ανάλυσης συναισθήματος

Η προσέγγιση βασισμένη σε λεξικό υπερτερεί έναντι της μηχανικής μάθησης για τους εξής λόγους:

- Η κατασκευή λεξικού είναι μια εύκολη διαδικασία και επιπλέον κατασκευάζοντας ένα λεξικό καλύπτουμε ένα ευρύ φάσμα λέξεων, ενώ οι τεχνικές ταξινόμησης είναι επίπονες και δεν εφαρμόζονται σε όλα τα πεδία.
- Η μεθοδολογία κατασκευής λεξικού προτιμάται επίσης στο πεδίο της γλωσσολογικής αξιολόγησης του κειμένου σε αντίθεση με τη μηχανική μάθηση.
- Οι τεχνικές που χρησιμοποιούν τα λεξικά δεν απαιτούν εκπαιδευμένα σύνολα δεδομένων. Αντίθετα τα συστήματα μηχανικής μάθησης στηρίζονται κατά βάση σε τέτοια σύνολα και για το λόγο αυτό αποφέρουν μη ακριβή αποτελέσματα, αφού είναι προσανατολισμένα σε πεδία.
- Επίσης η μηχανική μάθηση κάνει προβλέψεις για λέξεις που εμφανίζονται για πρώτη φορά στα σύνολα δεδομένων, γεγονός που μπορεί να μας οδηγήσει σε λανθασμένα αποτελέσματα.

1.6 Ταξινόμηση Συναισθήματος

Η ταξινόμηση συναισθήματος είναι η τεχνική αναγνώρισης και συναισθηματικής ομαδοποίησης των προτάσεων. Η ταξινόμηση αυτή χωρίζεται σε τρία (3) πεδία:

1) **Ο προσδιορισμός Συναισθηματικού προσανατολισμού με την ταξινόμηση της πολικότητας**: Σύμφωνα με τη μέθοδο της ταξινόμησης πολικότητας προσανατολίζονται οι απόψεις που εκφράζουν οι άνθρωποι σε θετικές, αρνητικές ή ουδέτερες. Η ταξινόμηση πολικότητας μπορεί να βελτιωθεί με την αφαίρεση των αντικειμενικών προτάσεων ενός κειμένου.
2) **Ο προσδιορισμός Υποκειμενικότητας**: Αυτή η τεχνική ίσως είναι πιο δύσκολη από τη ταξινόμηση της πολικότητας, καθώς ο προσανατολισμός των απόψεων σε θετικές, αρνητικές ή ουδέτερες γίνεται σύμφωνα με το περιεχόμενο ενός κειμένου.
3) **Ο προσδιορισμός Σθένους του Προσανατολισμού**: Η μέθοδος αυτή εξετάζει τη θετική ή αρνητική άποψη ενός κειμένου με κλίμακα διαβάθμισης. Υποδηλώνει την ένταση ενός συναισθήματος για το συγκεκριμένο κείμενο που εξετάζει.

1.7 Κατηγορίες Τεχνικών Ανάλυσης Συναισθήματος

Οι μέθοδοι για την ανάλυση συναισθήματος κατατάσσονται σε μία από τις παρακάτω κατηγορίες.

1.7.1 Keyword spotting

Οι μέθοδοι που ανήκουν σε αυτή την κατηγορία χρησιμοποιούν κάποιες ενδεικτικές λέξεις για την αναγνώριση συναισθήματος που ονομάζονται affect words. Η προσέγγιση αυτή θεωρείτε απλοϊκή γιατί βασίζεται μόνο σε λέξεις που προσδίδουν ξεκάθαρα κάποιο συναίσθημα με αποτέλεσμα να μην μπορεί να γίνει σωστή ανάλυση συναισθήματος όταν για παράδειγμα υπάρχει άρνηση μέσα στο κείμενο ή όταν το συναίσθημα εκφράζεται μέσα από λέξεις ή φράσεις που δεν κάνουν χρήση λέξεων.

1.7.2 Lexical affinity

Στην κατηγορία αυτή γίνεται χρήση των affect words όπως και στη παραπάνω κατηγορία με τη διαφορά πως αυτή τη φορά προσδίδεται σε κάθε λέξη μεγάλη πιθανότητα να σχετίζεται με κάποιο συναίσθημα. Για παράδειγμα η λέξη “accident” έχει 75% πιθανότητα να έχει αρνητική επίδραση. Οι μέθοδοι που ανήκουν σε αυτή τη κατηγορία έχουν καλύτερη επίδοση από τις μεθόδους που ανήκουν στην κατηγορία keyword spotting , αλλά υπάρχουν και περιπτώσεις που αποτυγχάνουν.

1.7.3 Statistical Methods

Αυτή η προσέγγιση κάνει χρήση αλγορίθμων μηχανικής μάθησης σε συνδυασμό με κείμενα που έχουν καταταχθεί χειροκίνητα σε κάποιο συναίσθημα με σκοπό να δημιουργηθεί μια μηχανή αναγνώρισης εγγράφων. Οι μέθοδοι αυτοί λειτουργούν καλά σε μεγάλα κείμενα.

1.7.4 Concept-based approaches

Οι μέθοδοι αυτές εστιάζουν στη σημασιολογική ανάλυση του κειμένου μέσω της χρήσης σημασιολογικών δικτύων που επιτρέπουν την ομαδοποίηση νοητικών και συγκινησιακών πληροφοριών που σχετίζονται με τις απόψεις της φυσικής γλώσσας. Στόχος των μεθόδων αυτών είναι να συναχθεί η σημασιολογική και συναισθηματική πληροφορία που σχετίζεται με τις απόψεις της φυσικής γλώσσας

Κεφάλαιο 2: Μηχανική μάθηση

2.1 Ορισμός

Η **Μηχανική Μάθηση** (Machine learning) είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Χρησιμοποιείται κυρίως για εύρεση και κατασκευή αλγορίθμων, οι οποίοι με την χρήση των κατάλληλων εισόδων μπορούν να κάνουν τις αντίστοιχες προβλέψεις και να εξάγουν αποτελέσματα μέσα σε ένα συγκεκριμένο όριο κατόπιν στατιστικής ανάλυσης. Επίσης συνδέεται αρκετά με την υπολογιστική στατιστική, η οποία ασχολείται με την πρόβλεψη αποτελεσμάτων μέσω της χρήσης των υπολογιστών καθώς και με την μαθηματική βελτιστοποίηση, η οποία με τη σειρά της παρέχει τις μεθόδους, τη θεωρία και τους τομείς εφαρμογής.

Η χρήση της έχει βοηθήσει σε μεγάλο βαθμό πολλούς γνώστες της τεχνολογίας, αφού η λειτουργία της ήταν και συνεχίζει να είναι ευρέως διαδεδομένη. Μερικά παραδείγματα στα οποία έχει προσφέρει σημαντικό όφελος είναι τα εξής:

- Η οπτική αναγνώριση χαρακτήρων (OCR), η οποία μετατρέπει σαρωμένες εικόνες κειμένων σε κείμενα κατανοητά και αναγνώσιμα από τον υπολογιστή.
- Η αναγνώριση προσώπου, όπου επιτρέπει στους χρήστες να μοιράζονται ένα σύνολο από ψηφιακές εικόνες και βίντεο.

- Τα φίλτρα spam, τα οποία φιλτράρουν τα εισερχόμενα e-mails και έτσι απορρίπτουν τα ανεπιθύμητα(spam e-mails).
- Οι μηχανισμοί συστάσεων, οι οποίοι προτείνουν ταινίες και τηλεοπτικές εκπομπές σύμφωνα με τα ενδιαφέροντα των χρηστών.
- Η αυτό-οδήγηση, όπου τα αυτοκίνητα βασίζουν την πλοήγησή τους σε υπολογιστικές μηχανές.

2.2 Είδη μηχανικής μάθησης

Στον τομέα της μηχανικής μάθησης αναπτύσσονται τρεις μέθοδοι μάθησης, σύμφωνα με τον τρόπο, τον οποίο δίνονται οι πληροφορίες στο σύστημα. Οι μέθοδοι αυτοί είναι:

1η Η **Εποπτευόμενη Μάθηση** (Supervised Learning) όπου ο υπολογιστής δέχεται δεδομένα εισόδου μαζί με τα αντίστοιχα επιθυμητά δεδομένα εξόδου. Ο στόχος της είναι, ο αλγόριθμος να μπορεί να μαθαίνει μέσα από την σύγκριση των πραγματικών εξόδων και με αυτών που δίνονται για να εντοπίσει τυχόν σφάλματα και παράλληλα να πραγματοποιήσει τις κατάλληλες αλλαγές σύμφωνα με το μοντέλο που δημιουργεί.

Χρησιμοποιείται σε προβλήματα Ταξινόμησης(Classification), Πρόγνωσης (Prediction) και Διερμηνείας (Interpretation).

2η Η **Μη Εποπτευόμενη Μάθηση** (UnSupervised Learning) όπου ο αλγόριθμος δημιουργεί ένα μοντέλο για να δέχεται ένα σύνολο δεδομένων με την μορφή παρατηρήσεων χωρίς όμως να γνωρίζει τις επιθυμητές εξόδους. Σκοπός είναι ο υπολογιστής να βρει μόνος του τον τρόπο επίλυσης του παραπάνω προβλήματος.

Χρησιμοποιείται σε προβλήματα Ανάλυσης Συσχετισμών (Association Analysis) και Ομαδοποίησης (Clustering).

3η Η **Ενισχυτική Μάθηση** (Reinforcement Learning) όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον.

Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ.

2.3 Μηχανική μάθηση και εξόρυξη δεδομένων

Συχνά όταν ακούμε για μηχανική μάθηση ακούμε και για εξόρυξη δεδομένων, πως όμως αυτά τα δύο συνδέονται μεταξύ τους?

Όπως προαναφέραμε με τη μηχανική μάθηση δημιουργούμε ή χρησιμοποιούμε αλγορίθμους ,που μας οδηγούν στην πρόβλεψη. Για να οδηγηθούμε στη πρόβλεψη αυτή, θα πρέπει να υπάρχουν δεδομένα και να εξάγονται από αυτά οι πληροφορίες που χρειαζόμαστε. Εδώ, έρχεται η **εξόρυξη δεδομένων** που δεν είναι τίποτα άλλο από τεχνικές * χρήσιμες για να αναλύσουμε πολύ μεγάλες συλλογές από δεδομένα και να εξάγουμε τις απαραίτητες πληροφορίες από αυτά.

Βασικοί αλγόριθμοι εξόρυξης δεδομένων είναι οι : C4.5, k-means,SVM (Support Vector Machine), a priori, EM (Expectation Maximization), PageRank, AdaBoost, kNN, Naive Bayes και CART, που καλύπτουν τα σημαντικότερα πεδία στην έρευνα της εξόρυξης δεδομένων , δηλαδή τα πεδία της ταξινόμησης, της ομαδοποίησης, της παλινδρόμησης, της στατιστικής μάθησης, της ανάλυσης συνδέσεων (association analysis) και της εξόρυξης συνδέσμων (link mining).

*τεχνικές=αλγόριθμοι

2.4 Προβλήματα της προσέγγισης με μηχανική μάθηση

Ένα από τα σημαντικότερα προβλήματα κατά τη χρήση της μηχανική μάθησης στην ανάλυση συναισθήματος είναι πως σε περίπτωση εισαγωγής προς ανάλυση κειμένου διαφορετικής δομής και νοηματικού τομέα βλέπουμε την απόδοση του συστήματος να πέφτει σημαντικά. Ωστόσο, υπάρχουν πολλοί τομείς για τους οποίους είναι δύσκολο να βρεθεί αρκετό υλικό εκπαίδευσης.

2.4.1 Μια άλλη κατηγοριοποίηση των προβλημάτων μηχανικής μάθησης

Α) Στην ταξινόμηση, όπου τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Αυτό συνήθως εμπίπτει στην επιτηρούμενη μάθηση. Τα φίλτρα Spam είναι ένα παράδειγμα ταξινόμησης, όπου οι είσοδοι είναι τα emails ή άλλα μηνύματα και οι κλάσεις είναι "spam" και "όχι spam".

Β)Στην παλινδρόμηση , επίσης πρόβλημα εποπτευόμενης μάθησης , τα αποτελέσματα είναι συνεχή και όχι διακριτά.

Γ) Στην συσταδοποίηση , ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτόν τον διαχωρισμό τυπική εργασία μη επιτηρούμενης μάθησης.

2.5 Λογισμικά μηχανικής μάθησης

Για τη πραγματοποίηση πειραμάτων μηχανικής μάθησης χρειαζόμαστε τη χρήση κάποιου λογισμικού. Τα λογισμικά που θα μπορούσαμε να επιλέξουμε αναλύονται παρακάτω.

2.5.1 Waikato Environment for Knowledge Analysis (Weka)

Το Weka είναι ένα δημοφιλής λογισμικό μηχανικής μάθησης γραμμένο σε Java και αναπτύχθηκε στο πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Είναι ελεύθερο λογισμικό υπό την άδεια GNU (General Public License = Γενική άδεια δημόσιας χρήσης), που είναι πιθανόν η περισσότερο δημοφιλής άδεια χρήσης ελεύθερου λογισμικού και είναι η άδεια που προστατεύει το μεγαλύτερο ποσοστό του ελεύθερου λογισμικού που υπάρχει μέχρι σήμερα.

Περιέχει μια συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και έχει πολλά πλεονεκτήματα. Μερικά από αυτά είναι:

1ο Έχει Δωρεάν διαθεσιμότητα υπό την GNU, όπως προαναφέραμε.

2ο πλεονέκτημα είναι η φορητότητα, διότι έχει υλοποιηθεί πλήρως στην γλώσσα προγραμματισμού Java και έτσι τρέχει σε σχεδόν κάθε σύγχρονη υπολογιστική πλατφόρμα.

3ο Διαθέτει μια ολοκληρωμένη συλλογή δεδομένων προ επεξεργασίας και τεχνικές μοντελοποίησης και

4ο Είναι εύκολο στη χρήση.

Επίσης το Weka υποστηρίζει διάφορες βασικές διεργασίες εξόρυξης δεδομένων, πιο συγκεκριμένα υποστηρίζει προ επεξεργασία δεδομένων, ομαδοποίηση, ταξινόμηση, παλινδρόμηση, απεικόνιση και δυνατότητα επιλογής. Ωστόσο δεν είναι ικανό για εξόρυξη από πολύ-σχεσιακές βάσεις δεδομένων και οι αλγόριθμοί του δεν καλύπτουν προς το παρόν τη μοντελοποίηση αλληλουχιών.

2.5.2 Matlab

Ένα άλλο λογισμικό μηχανικής μάθησης που θα μπορούσαμε να χρησιμοποιήσουμε είναι το MATLAB. Το MATLAB χρησιμοποιείτε συνήθως για σχεδίαση λειτουργιών και δεδομένων, υλοποίηση αλγορίθμων, δημιουργία διεπαφών χρήστη και διασύνδεση με προγράμματα γραμμένα σε άλλες γλώσσες , συμπεριλαμβανομένων C, C++, C#, Java και Python. Επιπλέον το MATLAB θα μπορούσε να χρησιμοποιηθεί και ως λογισμικό βαθιάς μάθησης. Παρόλο που το MATLAB προορίζεται κυρίως για αριθμητική υπολογιστική, μια προαιρετική εργαλειοθήκη χρησιμοποιεί

το συμβολικό μηχανισμό MuPAD , επιτρέποντας την πρόσβαση σε συμβολικές υπολογιστικές ικανότητες.

2.6 Εφαρμογές της μηχανικής μάθησης

Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση.

2.7 Μηχανική και βαθιά μάθηση

Τόσο η μηχανική όσο και η «βαθιά μάθηση» είναι υποσύνολα της τεχνητής νοημοσύνης, ωστόσο η τελευταία δεν είναι παρά η εξέλιξη της πρώτης. Στη μηχανική μάθηση, οι αλγόριθμοι που δημιουργούνται από τους ανθρώπους - προγραμματιστές είναι υπεύθυνοι για την ανάλυση των δεδομένων που λαμβάνουν και την εκμάθηση από αυτά. Κάπως έτσι καταλήγουν να λαμβάνουν αποφάσεις βάσει αυτών που μαθαίνουν από τα δεδομένα.

Η βαθιά εκμάθηση από την άλλη μαθαίνει μέσω ενός τεχνητού νευρικού δικτύου που λειτουργεί καθ' εικόνα και ομοίωση του ανθρώπινου εγκεφάλου επιτρέποντας στο μηχάνημα να αναλύει δεδομένα με μια συγκεκριμένη διαδικασία όπως πάνω κάτω κάνουν οι άνθρωποι. Δηλαδή οι μηχανές βαθιάς μάθησης δεν απαιτούν από έναν άνθρωπο - προγραμματιστή να τους πει τι να κάνουν με τα δεδομένα αυτά ώστε να αρχίσουν να μαθαίνουν από αυτά. Αυτό πραγματοποιείται από την εξαιρετική ποσότητα δεδομένων που συλλέγουν και καταναλώνουν οι ίδιοι - τα δεδομένα δηλαδή είναι το καύσιμο για την λειτουργία των μοντέλων βαθιάς μάθησης.

2.7.1 Που χρησιμοποιείται η βαθιά μάθηση

Μερικά παραδείγματα που χρησιμοποιείται η βαθιά μάθηση είναι:

- **Εμπειρία πελάτη**: Η μηχανική μάθηση χρησιμοποιείται ήδη από πολλές επιχειρήσεις για να βελτιώσει την εμπειρία των καταναλωτών. Ωστόσο υπάρχουν και παραδείγματα βαθιάς μάθησης που χρησιμοποιούνται για την εξυπηρέτηση του πελάτη, όπως τα chatbots ορισμένων καταστημάτων που σας λύνουν ανά πάσα ώρα και στιγμή τυχόν απορίες για προϊόντα και αγορές. Καθώς δε, η βαθιά μάθηση ωριμάζει αναμένουμε ότι ο κλάδος του εμπορίου θα είναι ένας από τους βασικούς που θα χρησιμοποιηθεί.

- **Μεταφράσεις:** Παρότι το μοντέλο της αυτόματης μετάφρασης δεν είναι καινούργιο, η βαθιά μάθηση συμβάλλει στη βελτίωση της εμπειρίας αυτόματης μετάφρασης κειμένου χρησιμοποιώντας σταυροειδή δίκτυα νευρωνικών δικτύων και να επιτρέψουν μεταφράσεις από εικόνες.
- **Αναγνώριση γλώσσας ομιλούντος:** Η βαθιά μάθηση ξεκινά να αναγνωρίζει τις διαλέκτους μας γλώσσας, χωρίς την ανάμειξη ανθρώπου.
- **Αυτόνομη οδήγηση:** Κάποια συστήματα βαθιάς μάθησης εκπαιδεύονται να αναγνωρίζουν πινακίδες του δρόμου ενώ άλλοι εκπαιδεύονται να αναγνωρίζουν τους πεζούς. Καθώς ένα αυτοκίνητο κινείται στους δρόμους, μπορεί να χρησιμοποιεί εκατομμύρια αλγορίθμους τεχνητής νοημοσύνης που το βοηθούν να οδηγεί και να αλληλοεπιδρά με το εξωτερικό περιβάλλον.

Κεφάλαιο 3: Αλγόριθμοι

3.1Ορισμός

Ως αλγόριθμο ορίζουμε μια σειρά από εντολές που έχουν αρχή και τέλος, είναι σαφείς και έχουν ως σκοπό την επίλυση κάποιου προβλήματος. Η λέξη αλγόριθμος έχει καθιερωθεί με την έννοια «συστηματική διαδικασία αριθμητικών χειρισμών». Τη σημερινή της σημασία την οφείλει στη γρήγορη ανάπτυξη των ηλεκτρονικών υπολογιστών στα μέσα του 20ου αιώνα.

3.2 Αλγόριθμοι κατηγοριοποίησης

Η κατηγοριοποίηση αποτελεί μια από τις βασικές εργασίες στην εξόρυξη δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου το όποιο, με βάση τα χαρακτηριστικά αυτά, αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από:

- έναν καλά καθορισμένο ορισμό των κατηγοριών(κλάσεων)
- το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου

Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί.

Το παρακάτω σχήμα (γράφημα 3.2.Α) εμφανίζει συνοπτικά την διαδικασία της κατηγοριοποίησης, κατά την οποία το ,μοντέλο δέχεται ως είσοδο ένα σύνολο

χαρακτηριστικών, από δεδομένα με γνωστή κλάση, για να προβλέψει την ετικέτα των μη ταξινομημένων δεδομένων.

Κάθε τεχνική κατηγοριοποίησης βασίζεται σε έναν αλγόριθμο μάθησης, για να βρει το μοντέλο που περιγράφει καλύτερα τη σχέση

Γράφημα 3.2.Α: Η διαδικασία της κατηγοριοποίησης

Το μοντέλο, πρέπει αφενός να ταιριάζει (fit)καλά στα δεδομένα εισόδου και αφετέρου να προβλέπει σωστά τις κλάσεις των μη ταξινομημένων εγγραφών.

Στην επιβλεπόμενη μάθηση (supervised learning), χρησιμοποιείται ένα σύνολο εκπαίδευσης (training set), που αποτελείται από εγγραφές με γνωστές ετικέτες. Το μοντέλο κατηγοριοποίησης χρησιμοποιεί το σύνολο εκπαίδευσης, ώστε με τη σειρά του να εκπαιδευτεί και να προβλέψει σωστά τις κλάσεις μη ταξινομημένων εγγράφων, που ανήκουν στο σύνολο ελέγχου (test set).Η αποτελεσματικότητα του μοντέλου κατηγοριοποίησης θα κριθεί από το πλήθος των εγγραφών του συνόλου ελέγχου, για τις οποίες έγινε σωστή πρόβλεψη της κλάσης και αυτήν για τις οποίες έγινε η πρόβλεψη των λαθών. Το πλήθος των σωστών και των λαθών προβλέψεων τοποθετείται σε έναν πίνακα που είναι γνωστός ως μέτρα σύγχυσης (confusion matrix)Το γράφημα 3.2.Β παρουσιάζει το πλήθος των σωστά και λάθος ταξινομημένων θετικών και αρνητικών εγγραφών όπου οι στήλες αντιπροσωπεύουν τις σωστές και λάθος προβλέψεις κάθε κλάσης και οι γραμμές το πλήθος των πραγματικών παραδειγμάτων που ανήκει σε κάθε κλάση.

	Πρόβλεψη	
Πραγματική Ετικέτα	**Θετική**	**Αρνητική**
Θετική	**TP** (True Positives)	**FP** (False Positives)
Αρνητική	**FN** (False Negatives)	**TN** (True Negatives)

Γράφημα 3.2.Β:Μέτρα Σύγχυσης

όπου,

TP (true positives), είναι τα παραδείγματα που κατηγοριοποιήθηκαν σωστά ως θετικά

FP(false positives), είναι τα παραδείγματα που κατηγοριοποιήθηκαν λάθος ως θετικά

TN (true negatives), είναι τα παραδείγματα που κατηγοριοποιήθηκαν σωστά ως αρνητικά

FN (false negatives), είναι τα παραδείγματα που κατηγοριοποιήθηκαν λάθος ως αρνητικά.

3.2.1 Naive Bayes

Ο κατηγοριοποιητής Bayes είναι ένας από τους πιο δημοφιλής αλγόριθμους κατηγοριοποίησης κειμένου. Χρησιμοποιεί το γνωστό θεώρημα Bayes για να προβλέψει την κλάση αυτή που μεγιστοποιεί την μεταγενέστερη πιθανότητα. Ο κύριος στόχος είναι να εκτιμηθεί η συνδυαστική συνάρτηση πυκνότητας πιθανότητας για κάθε κλάση που μοντελοποιείται μέσω μιας πολυμεταβλητής κανονικής διανομής. Ο απλουστευμένος κατηγοριοποιητής Bayes υποθέτει ότι τα γνωρίσματα ,προς μελέτη, είναι ανεξάρτητα μεταξύ τους, παρόλα αυτά χρησιμοποιείται πολύ συχνά σε εφαρμογές. Για να χρησιμοποιηθεί όμως σε εφαρμογές θα πρέπει πρώτα να εκπαιδευτεί, έτσι ώστε να αρχικοποιηθούν οι παράμετροί του. Η εκπαίδευση γίνετε με τα εκπαιδευτικά έγγραφα, δηλαδή έγγραφα τα οποία περιέχουν και την κατηγορία την οποία ανήκουν. Μετά το τέλος της εκπαίδευσης ο Naive Bayes είναι σε θέση να υπολογίσει την πιθανή κατηγορία στην οποία ανήκει ένα νέο άγνωστο έγγραφο. Ένα από τα πλεονεκτήματα της χρήσης αυτού του μοντέλου είναι η απλότητα του.

Ας δούμε πως δουλεύει ο αλγόριθμος αυτός.

Υποθέτουμε ότι έχουμε ένα σετ από δεδομένα D που αποτελείται από n σημεία xi σε ένα d-διάστατο χώρο και έστω yi η κλάση που ανήκει κάθε σημείο. k}.Ο κατηγοριοποιητής Bayes απευθείας χρησιμοποιεί το θεώρημα Bayes για να προβλέψει την κλάση για το καινούργιο στιγμιότυπο x. Εκτιμά την μεταγενέστερη πιθανότητα P(ci | x) για κάθε κλάση ci και διαλέγει την κλάση με τη μεγαλύτερη πιθανότητα. Με άλλα λόγια η προβλεπόμενη κλάση ουσιαστικά εξαρτάται από την πιθανότητα της ίδιας της κλάσης λαμβάνοντας υπ’ όψιν και την προγενέστερη

πιθανότητα που είχε. Για να κατηγοριοποιήσουμε τα σημεία θα πρέπει να εκτιμηθούν οι δύο όροι που χρειάζονται κατευθείαν από το σετ των δεδομένων D. Σύμφωνα με την στατιστική ,η από κοινού πιθανότητα ορίζεται ως:

$$\mathbf{P(X \cap Y) = P(Y|X) * P(X) = P(X|Y^*)P(Y)}$$

[Σχέση 2.4]

Μετασχηματίζοντας τη σχέση (2.4), καταλήγουμε στην ακόλουθη σχέση:

$$P(Y|X) = (P(X|Y) * P(Y))/P(X)$$

[Σχέση 2.5]

που είναι το θεώρημα του Bayes και χρησιμοποιείται για την πραγματοποίηση των προβλέψεων, στις οποίες αναφερθήκαμε νωρίτερα.

Η πιθανότητα P(Y)είναι η εκ των προτέρων πιθανότητα (prior pro-bability) και η πιθανότητα P(Y|X) είναι η εκ των υστέρων πιθανότητα.

Λόγω της απλότητας του και της μικρής πολυπλοκότητας του ο ταξινομητής Naive Bayes προτιμάται όταν υπάρχουν περιορισμένοι διαθέσιμοι υπολογιστικοί πόροι καθώς και σε περιπτώσεις στις οποίες επιθυμούμε τη γρήγορη εκπαίδευση του συστήματος.

3.2.2 Μηχανές Διανυσμάτων Υποστήριξης - Support Vector Machines (SVM)

Η μέθοδος των ΜΔΣ χρησιμοποιείται για κατηγοριοποίηση γραμμικών και μη γραμμικών δεδομένων. Με κατάλληλη μη γραμμική αντιστοίχιση σε μια επαρκώς υψηλή διάσταση τα δεδομένα που ανήκουν σε δύο διαφορετικές κλάσεις μπορούν πάντα να διαχωριστούν με ένα υπερ-επίπεδο. Η μέθοδος των ΜΔΣ βρίσκει αυτό το υπερεπίπεδο χρησιμοποιώντας διανύσματα στήριξης, ουσιώδεις δηλαδή πλειάδες προς εκπαίδευση, και περιθώρια, που ορίζονται από τα διανύσματα στήριξης. Παρόλο που ο χρόνος εκπαίδευσης ακόμα και της πιο γρήγορης μεθόδου ΜΔΣ μπορεί να είναι εξαιρετικά αργός, τα αποτελέσματα είναι πολύ ακριβή, διότι στηρίζονται στη δυνατότητα να μοντελοποιούν πολύπλοκα μη γραμμικά όρια απόφασης. Οι ΜΔΣ χρησιμοποιούνται τόσο για προβλήματα κατηγοριοποίησης όσο και για προβλήματα αριθμητικών προβλέψεων. Έχουν βρει εφαρμογή σε πολλούς

τομείς όπως την αναγνώριση χαρακτήρων γραπτού λόγου, αναγνώριση αντικειμένων, αναγνώριση φωνής και πρόβλεψη χρονοσειρών.

Στην ουσία η βασική ιδέα των ΜΔΣ είναι να δημιουργήσει ένα υπερ-επίπεδο ως την επιφάνεια απόφασης με τέτοιο τρόπο ώστε η απόσταση ανάμεσα στα θετικά και στα αρνητικά παραδείγματα να είναι το μέγιστο. Η μηχανή επιτυγχάνει αυτή την επιθυμητή ιδιότητα, ακολουθώντας μια αρχή που έχει ως βάση την θεωρία στατιστικής μάθησης.

Σαν πρώτο βήμα είναι η εκπαίδευση της μηχανής. Αρχικά αντιστοιχεί τα πρώτα δεδομένα σε έναν χώρο υψηλών διαστάσεων και με βάση αυτήν την διάσταση, ψάχνει για γραμμικά διαχωριζόμενα υπερ-επίπεδα, δηλαδή γραμμές που χωρίζουν τις δύο κλάσεις, αλλά δεν είναι όλες βέλτιστες. Συνεπώς , στόχος είναι η εύρεση του βέλτιστου υπέρ-επίπεδου δηλαδή αυτού που ελαχιστοποιεί το σφάλμα κατηγοριοποίησης στα άγνωστα δεδομένα. Στο δεύτερο βήμα και από τη στιγμή που θα βρεθεί το ελάχιστο για ένα δεδομένο σύνολο δεδομένων, η δοσμένη ΜΔΣ θα συγκλίνει πάντα ντετερμινιστικά στην ίδια λύση. Μετά την εκπαίδευση η ΜΔΣ μπορεί πλέον να αναθέτει νέα στοιχεία σε κάθε κατηγορία.

Στην περίπτωση που τα δεδομένα δεν διαχωρίζονται με γραμμικό τρόπο δεν μπορούν να εφαρμοστούν όσα προαναφέρθηκαν. Αφού λοιπόν δε μπορούμε να τραβήξουμε κάποια ευθεία γραμμή ώστε να διαχωριστούν οι κλάσεις θα πρέπει να επεκτείνουμε την προσέγγισή μας και να δημιουργήσουμε μη γραμμικές ΜΔΣ. Τέτοιες ΜΔΣ είναι ικανές να βρίσκουν μη γραμμικά όρια αποφάσεων, όπως για παράδειγμα μη γραμμικές υπερ-επιφάνειες, στο χώρο εισόδου. Η μέθοδος επέκτασης περιλαμβάνει 2 κύρια βήματα. Το πρώτο βήμα περιλαμβάνει το μετασχηματισμό των αρχικών δεδομένων εισόδων σε υψηλότερης διάστασης χώρο χρησιμοποιώντας μη γραμμική αντιστοίχιση. Μόλις τα δεδομένα μετασχηματιστούν κατά το δεύτερο βήμα αναζητάμε για ένα γραμμικώς διαχωρίσιμο υπερεπίπεδο στο νέο χώρο. Τελικά καταλήγουμε με ένα τετραγωνικό πρόβλημα βελτιστοποίησης που μπορεί να επιλυθεί με χρήση γραμμικών ΜΔΣ. Το υπερ-επίπεδο μέγιστου περιθωρίου που βρέθηκε στον καινούργιο χώρο αντιστοιχεί στη μη γραμμική διαχωριστική υπερ-επιφάνεια στον πραγματικό αρχικό χώρο. Τελικά στόχος του ταξινομητή είναι να διαλέξει το υπερ-επίπεδο που αντιπροσωπεύει το ορθότερο όριο απόφασης για το σύνολο ελέγχου, μεγιστοποιώντας τα όρια απόφασης τόσο όσο χρειάζεται για να ελαχιστοποιηθούν τα σφάλματα γενίκευσης.

3.2.3 Κατηγοριοποίηση με βάση τους k εγγύτερους γείτονες (k-Nearest Neighbors)

Τα μοντέλα κατηγοριοποίησης που έχουμε περιγράψει μέχρι στιγμής είναι όλα μοντέλα πρόθυμης όπως λέγεται μάθησης. Αυτό σημαίνει ότι όταν δίνεται ένα

σύνολο πλειάδων για εκπαίδευση θα κατασκευαστεί ένα γενικευμένο μοντέλο προτού γίνουν δεκτές νέες πλειάδες προς κατηγοριοποίηση. Μπορούμε να θεωρήσουμε ότι το μοντέλο είναι έτοιμο και πρόθυμο να κατηγοριοποιήσει τις άγνωστες μέχρι πριν πλειάδες. Αντιθέτως στην οκνηρή προσέγγιση περιμένουμε μέχρι την τελευταία στιγμή ,προτού γίνει κατασκευή οποιουδήποτε μοντέλου, να κατηγοριοποιηθεί το σύνολο των δοκιμαστικών πλειάδων. Ένας οκνηρός λοιπόν μαθητευόμενος με δεδομένη μια εκπαιδευόμενη πλειάδα απλά την αποθηκεύει, ή κάνει μια πολύ μικρή επεξεργασία της, και περιμένει μέχρι να λάβει μια δοκιμαστική πλειάδα. Μόνο μόλις δει τη δοκιμαστική πλειάδα εφαρμόζει γενίκευση για να κατηγοριοποιηθεί με βάση την ομοιότητά της με τις αποθηκευμένες εκπαιδευμένες πλειάδες. Ανόμοια λοιπόν με τις πρόθυμες μεθόδους μάθησης, οι οκνηροί μαθητευόμενοι κάνουν λιγότερη δουλειά όταν εμφανίζεται μια πλειάδα προς εκπαίδευση και περισσότερη όταν είναι να γίνει κατηγοριοποίηση ή αριθμητική πρόβλεψη. Επειδή λοιπόν οι οκνηροί μαθητευόμενοι αποθηκεύουν τις πλειάδες ή αλλιώς στιγμιότυπα αναφέρονται και ως μαθητευόμενοι βασισμένοι σε στιγμιότυπα, διότι σε αυτά βασίζεται όλη η διαδικασία της μάθησης. Για να γίνει μια κατηγοριοποίηση ή μια αριθμητική πρόβλεψη η χρήση οκνηρών μαθητευόμενων μπορεί να είναι υπολογιστικά ακριβή. Απαιτούν αποδοτικές τεχνικές αποθήκευσης και θα πρέπει να προσαρμόζονται σε εφαρμογές με υλικό που λειτουργεί με παράλληλο ποίηση, ενώ προσφέρουν πολύ λίγη εξήγηση για την εσωτερική δομή των δεδομένων. Παρόλα αυτά όμως υποστηρίζουν από τη φύση τους τη στοιχειώδη εκπαίδευση. Μπορούν να μοντελοποιήσουν πολύπλοκους χώρους αποφάσεων που έχουν υπερ-πολυγωνικά σχήματα ,τα οποία δεν είναι εύκολα περιγράψιμα από άλλους αλγόριθμους. Όταν τα δεδομένα μας είναι ονομαστικά ή κατηγοριακά μια απλή μέθοδος είναι να συγκρίνουμε απευθείας τις τιμές του χαρακτηριστικού που ζητάμε της πλειάδας Χ1 με αυτήν της πλειάδας Χ2. Αν αυτές είναι ίδιες τότε η διαφορά μεταξύ των δύο είναι 0 αλλιώς αν είναι διαφορετικές ισούται με 1. Γενικά αν η τιμή ενός δεδομένου χαρακτηριστικού Α λείπει από μια πλειάδα Χ1 ή/και από μια πλειάδα Χ2 υποθέτουμε τη μέγιστη πιθανή διαφορά. Αν το Α είναι αριθμητικό και λείπει κι από τις δύο πλειάδες τότε ξανά η διαφορά ισούται με 1. Αν όμως μόνο μία από τις τιμές λείπει και η άλλη είναι παρούσα και κανονικοποιημένη (ν') τότε η διαφορά ισούται είτε με |1-ν'| είτε |0-ν'|, όποια από αυτές είναι μεγαλύτερη. Ο αριθμός k των κοντινότερων γειτόνων προσδιορίζεται πειραματικά. Ξεκινάμε με k=1 και χρησιμοποιούμε ένα δοκιμαστικό σύνολο για την εκτίμηση του ποσοστού σφάλματος του κατηγοριοποιητή. Η διαδικασία επαναλαμβάνεται κάθε φορά αυξάνοντας το k επιτρέποντας έναν ακόμα γείτονα. Τελικά επιλέγεται αυτή η τιμή που δίνει το μικρότερο ποσοστό σφάλματος. Γενικά όσο μεγαλύτερος είναι ο αριθμός των δοκιμαστικών πλειάδων τόσο μεγαλύτερο θα είναι το k. Όσο ο αριθμός των πλειάδων τείνει στο άπειρο και k=1 το ποσοστό σφάλματος δε μπορεί να είναι χειρότερο από το διπλάσιο του ποσοστού

σφάλματος Bayes, όπου το τελευταίο είναι το θεωρητικό ελάχιστο. Εάν επιπλέον και το k τείνει στο άπειρο τότε το ποσοστό σφάλματος πλησιάζει αυτό του Bayes.

Ένα πρακτικό ζήτημα που εφαρμόζεται στον αλγόριθμο είναι πως η απόσταση μεταξύ των δειγμάτων υπολογίζεται βασιζόμενη σε όλα τα χαρακτηριστικά του δείγματος, σε αντίθεση με μεθόδους που επιλέγουν μόνο ένα υποσύνολο των χαρακτηριστικών. Κατά συνέπεια η μετρική ομοιότητα που χρησιμοποιείται στους k-εγγύτερους γείτονες εξαρτάται κι από άσχετα χαρακτηριστικά οπότε μπορεί να είναι παραπλανητική. Η δυσκολία λοιπόν που προκύπτει λόγω του ότι αυτά τα χαρακτηριστικά είναι παρόντα αποκαλείται “κατάρα της διάστασης” και η τεχνική των εγγύτερων γειτόνων είναι ευαίσθητη σε αυτό το πρόβλημα. Όσον αφορά τη χρονική πολυπλοκότητα ο αλγόριθμος των k-εγγύτερων γειτόνων μπορεί να είναι εξαιρετικά αργός όταν κατηγοριοποιεί δοκιμαστικές πλειάδες.

3.3 Αλγόριθμοι Παλινδρόμησης

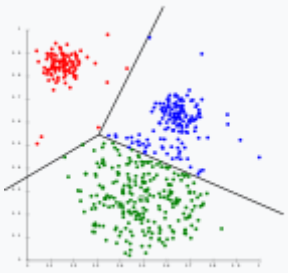
Οι αλγόριθμοι παλινδρόμησης χρησιμοποιούνται για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Σκοπός τους είναι η εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, αλλιώς καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής. Οι αλγόριθμοι παλινδρόμησης προϋποθέτουν ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο ς ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

3.4 Αλγόριθμοι συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης αυτό που πραγματικά κάνουν είναι να παίρνουν ένα σύνολο από «αντικείμενα» και να τα διαχωρίζουν σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση αντικειμένων σε ίδια ομάδα μεταφράζεται ως ομοιότητα των αντικειμένων , ενώ αντικείμενα που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια.

3.4.1 K-means

Ο συγκεκριμένος αλγόριθμος είναι από τους πιο πολυεφαρμοσμένους και είναι η ρίζα για πολλούς άλλους. Ανήκει στην κατηγορία της επίπεδης συσταδοποίησης διότι παράγει ένα σύνολο συσταδοποιήσεων οι οποίες δεν έχουν κάποια ιδιαίτερη δομή μεταξύ τους. Ο αλγόριθμος έχει ως στόχο τη βελτιστοποίηση μίας συνάρτησης – της συνάρτησης κόστους.



Γράφημα 3.4.1: K-means

Αρχικά έχουμε k-ομάδες, με την κάθε ομάδα *να αντιπροσωπεύεται από το μέσο διάνυσμα*. Ο αλγόριθμος λειτουργεί σε δύο βήματα:

1. Κατά τη φάση της διαμέρισης γίνεται προσπέλαση των διανυσμάτων και για κάθε διάνυσμα βρίσκουμε την απόστασή του από τις υπάρχουσες ομάδες. Η απόσταση μεταξύ ενός διανύσματος και μίας ομάδας είναι η ευκλείδεια απόσταση από το μέσο διάνυσμα. Για να υπολογίσουμε τις αποστάσεις, προσδιορίζονται N αριθμοί και σχηματίζονται k-σύνολα διανυσμάτων , ένα σύνολο για κάθε ομάδα.
2. Κατά τη φάση της ενημέρωσης, γίνονται οι κατάλληλες τροποποιήσεις στα μέσα διανύσματα, δηλαδή υπολογίζονται ξανά τα μέσα διανύσματα m_i για i=1,2...κ. Στον υπολογισμό του νέου m_i συνεισφέρει το αντίστοιχο σύνολο διανυσμάτων που υπολογίστηκε κατά το παραπάνω βήμα.

Ο αλγόριθμος ολοκληρώνεται όταν οι ενημερώσεις που γίνονται σε κάθε m_i είναι αμελητέες. Σημαντικό σημείο του αλγορίθμου είναι η αρχικοποίηση των k-διανυσμάτων.
Η χρονική πολυπλοκότητα του αλγόριθμου είναι O(Nkq) όπου q ο αριθμός των επαναλήψεων που πρέπει να εκτελέσει ο αλγόριθμος για να τερματίσει.

3.4.2 Ιεραρχικοί

Οι ιεραρχικοί αλγόριθμοι παράγουν μια ιεραρχία εμφωλιασμένων συσταδοποιήσεων. Οι αλγόριθμοι αυτοί χωρίζονται σε 2 υποκατηγορίες. Στους συσσωρευτικούς και στους διαιρετικούς

Οι συσσωρευτικοί ξεκινάνε με n ομάδες. Σε κάθε βήμα του αλγορίθμου b, το πλήθος των ομάδων μειώνεται κατά ένα συγχωνεύοντας δύο σε μία ομάδα έως φτάσουμε σε μία μοναδική που να εμπεριέχει όλα τα διανύσματα. Η εξέλιξη του αλγορίθμου μπορεί να αναπαρασταθεί γραφικά με ένα δενδρόγραμμα ανομοιότητας. Στο βήμα b, συγχωνεύονται οι ομάδες οι οποίες είναι το λιγότερο ανόμοιες.

Οι διαιρετικοί αλγόριθμοι ακολουθούν αντίστροφη διαδικασία. Ξεκινούν από μία ομάδα η οποία εμπεριέχει όλα τα διανύσματα και σε κάθε βήμα μία ομάδα

διασπάται σε δύο μέχρι να καταλήξουμε σε N ομάδες. Η πολυπλοκότητα τους είναι μεγαλύτερη από τους συσσωρευτικούς αφού η διάσπαση μίας ομάδας σε δύο μπορεί να γίνει κατά 2^{N-1}-1 τρόπους και η επιλογή της βέλτιστης διάσπασης πρακτικά είναι αδύνατη ακόμα και για μικρό N. Στην πράξη αυτό που γίνεται είναι ότι ο αλγόριθμος σε κάθε βήμα διασπά μία ομάδα αλλά όχι κατά βέλτιστο τρόπο.

Κεφάλαιο 4: Επεξεργασία Φυσικής Γλώσσας (nlp)

4.1 Ορισμός

Η **επεξεργασία φυσικής γλώσσας** (nlp) είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων (φυσικών) γλωσσών. Κατά συνέπεια, η nlp συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή.

4.2 Οι προκλήσεις της Φυσικής Γλώσσας (nlp)

Προκλήσεις της nlp περιλαμβάνουν την κατανόηση φυσικής γλώσσας, δηλαδή την προσπάθεια να καταστούν ικανοί οι υπολογιστές να εξάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα, αλλά και την παραγωγή φυσικής γλώσσας.

4.3 Σημαντικά πεδία έρευνας στην Επεξεργασία Φυσικής Γλώσσας

Στην φυσική επεξεργασία γλώσσας υπάρχουν αρκετά πεδία έρευνας. Μερικά από αυτά έχουν εφαρμογές στην καθημερινή ζωή, ενώ άλλα υφίστανται ως υποπεδία που υποβοηθούν την επίλυση μεγαλύτερων ζητημάτων. Το κριτήριο που ξεχωρίζει τα συχνότερα πεδία έρευνας από άλλα πιθανά και πραγματικά πεδία της nlp είναι το γεγονός ότι για το κάθε ένα από αυτά υπάρχει επίσημα ένας καλά ορισμένος χώρος εργασιών και επίλυσης ζητημάτων. Από τα πιο συχνά πεδία έρευνας στην επεξεργασία φυσικής γλώσσας είναι:

- **Η Ανάλυση λόγου** που είναι ένα σύνολο από μελέτες. Μία μελέτη αναφέρεται στην αναγνώριση της δομής του λόγου ενός συνδεόμενου κειμένου ενώ μια άλλη πιθανή μελέτη είναι η αναγνώριση και η κατηγοριοποίηση των γλωσσικών πράξεων σε ένα κομμάτι κειμένου.
- **Η Αυτόματη αναγνώριση ομιλίας** που στην ουσία είναι η αυτόματη μετατροπή του προ φερόμενου ανθρώπινου λόγου σε κείμενο από τους υπολογιστές.

- **Η Αυτόματη ερωταπόκριση,** δηλαδή η αναζήτηση της σωστής απάντησης σε μία δεδομένη ερώτηση διατυπωμένη σε ανθρώπινη γλώσσα.
- **Η Αυτόματη περίληψη.** Πρόκειται για τη παραγωγή μίας αναγνώσιμης περίληψης ενός κειμένου. Συχνά χρησιμοποιείτε για να παρέχει περιλήψεις σε κείμενα γνωστής διάταξης, όπως άρθρα στο οικονομικό μέρος μίας εφημερίδας.
- **Η Εξαγωγή πληροφοριών** που είναι η ανάκτηση πληροφοριών από μη δομημένα ή ημιδομημένα δεδομένα.
- **Η Επισήμανση των μερών του λόγου.** Στο πεδίο αυτό επισημαίνονται τα μέρη του λόγου σε μία δεδομένη πρόταση **.**
- **Η Κατανόηση φυσικής γλώσσας.** Εδώ γίνετε η μετατροπή κομματιών κειμένου σε πιο τυπικές αναπαραστάσεις όπως σε δομές λογικής πρώτου βαθμού, οι οποίες μπορούν να μεταχειριστούν ευκολότερα από τους υπολογιστές.
- **Η Μηχανική μετάφραση** που εδώ γίνεται αυτόματη μετάφραση ενός κειμένου από μία ανθρώπινη γλώσσα σε μία άλλη.
- **Η Σύνθεση ομιλίας** δηλαδή η αυτόματη ,τεχνητή παραγωγή του ανθρώπινου λόγου από τους υπολογιστές .

4.4 Τα προβλήματα της Επεξεργασίας φυσικής Γλώσσας (nlp)

Τα βασικότερα προβλήματα στην Επεξεργασία Φυσικής Γλώσσας είναι:

Α) η μη τήρηση γραμματικών κανόνων και

Β) η μη αναγνώριση λέξεων. Βασικές αιτίες για το πρόβλημα της μη αναγνώρισης λέξεων μπορεί να είναι, είτε να μην υπάρχουν οι λέξεις στο λεξικό, είτε να μη πρόκειται για λέξεις , είτε να έχουν καταχωρηθεί λανθασμένα.

Κεφάλαιο 5: Κοινωνικά δίκτυα

5.1 Ορισμός

Τα κοινωνικά δίκτυα είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Ο όρος σήμερα χρησιμοποιείται επίσης για να περιγράψει ιστοσελίδες οι οποίες επιτρέπουν την διεπαφή ανάμεσα στους χρήστες, πχ. με σχόλια, φωτογραφίες, άλλες πληροφορίες από σχετική βιβλιογραφία. Οι πιο γνωστές από αυτές τις ιστοσελίδες είναι το Facebook, Twitter, Instagram και Linkedin.

Ένα κοινωνικό δίκτυο είναι μια κοινωνική δομή που αποτελείται από ένα σύνολο παραγόντων, όπως άτομα ή οργανισμούς. Στο διαδίκτυο, τα κοινωνικά δίκτυα είναι

μία πλατφόρμα που συντηρείται για την δημιουργία κοινωνικών σχέσεων μεταξύ των ανθρώπων, που συνήθως αποτελούν ενεργά μέλη του κοινωνικού δικτύου, με κοινά ενδιαφέροντα ή δραστηριότητες.

5.2 Facebook

Το **Facebook** είναι μια πλατφόρμα κοινωνικής δικτύωσης που ξεκίνησε στις 4 Φεβρουαρίου του 2004. Οι χρήστες της μπορούν να επικοινωνούν μεταξύ τους μέσω μηνυμάτων με τις επαφές τους και να τους ειδοποιούν όταν ανανεώνουν τις προσωπικές πληροφορίες τους.

Επίσης ήταν και εξακολουθεί να είναι το πιο δημοφιλές κοινωνικό δίκτυο. Ο αριθμός των χρηστών είναι αρκετά μεγάλος και ανέρχεται στα 1.50 δισεκατομμύρια χρήστες ενεργούς κάθε μήνα.

5.2 Λογότυπο του μέσω κοινωνικής δικτύωσης Facebook.

5.3Twitter

Το Twitter είναι ένας ιστοχώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (μέχρι 280 χαρακτήρες), τα οποία ονομάζονται τουίτς (tweets). Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν κείμενα. Δημιουργήθηκε στις 21 Μαρτίου του 2006 από τον Τζακ Ντόρσεϊ και δημοσιεύθηκε τον Ιούλιο του ίδιου χρόνου. Η υπηρεσία έγινε γρήγορα δημοφιλής και είχει 305 εκατομμύρια ενεργούς χρήστες το 2015. Τα τελευταία χρόνια το twitter έχει κάνει μεγάλη στροφή και κινείται ραγδαία προς το δρόμο της επιτυχίας του facebook.

5.3 Λογότυπο του μέσω κοινωνικής δικτύωσης Twitter.

5.4Instagram

Το Instagram είναι μια δωρεάν εφαρμογή κοινωνικής δικτύωσης που δίνει την δυνατότητα επεξεργασίας και κοινοποίησης φωτογραφιών και βίντεο στο διαδίκτυο. Οι χρήστες μπορούν να μοιράζονται φωτογραφίες και βίντεο με τους ακολούθους τους (followers) ή με επιλεγμένη ομάδα φίλων, να σχολιάζουν και να δηλώνουν ότι μια δημοσίευση τους αρέσει. Δημιουργήθηκε από δύο απόφοιτους του Πανεπιστημίου του Στάντφορντ, τους Κέβιν Σίστρομ και Μάικ Κρίγκερ και ξεκίνησε τον Οκτώβριο του 2010. Μόλις δύο μήνες αργότερα, τον Δεκέμβριο του 2010, ο αριθμός των εγγεγραμμένων χρηστών έφτασε το 1.000.000. Σήμερα η εφαρμογή μετράει 20 δισεκατομμύρια φωτογραφίες από όλο τον κόσμο και 1 δισεκατομμύριο ενεργούς χρήστες. Το όνομα της προέρχεται από τον συνδυασμό της λέξης **Insta**nt (στιγμιαίο) και tele**gram** (τηλεγράφημα). Το 2012 η εφαρμογή αγοράστηκε από το Facebook, προς ένα 1 δισεκατομμύρια δολάρια.

5.4 Λογότυπο του μέσω κοινωνικής δικτύωσης Instagram.

5.5LinkedIn

Το LinkedIn είναι ιστοχώρος επαγγελματικής κοινωνικής δικτύωσης. Ιδρύθηκε τον Δεκέμβριο του 2002 από τον Ρέιντ Χόφμαν, αλλά ξεκίνησε επίσημα στις 5 Μαΐου του 2003. Η έδρα της εταιρίας είναι στη Σίλικον Βάλλεϋ και έχει γραφεία σε όλο τον κόσμο.[1] Τα εγγεγραμμένα μέλη του έχουν τη δυνατότητα να δημιουργήσουν το προσωπικό επαγγελματικό τους προφίλ, να συνδεθούν με άλλους χρήστες, να αναζητήσουν εργασία, αλλά και να δημιουργήσουν πελατολόγιο. Ο ιστοχώρος είναι διαθέσιμος σε 24 γλώσσες. Σήμερα θεωρείται ο πιο επιτυχημένος ιστοχώρος επαγγγελματικής κοινωνικής δικτύωσης παγκοσμίως, μετρώντας περισσότερους από 300 εκατομμύρια εγγεγραμμένους χρήστες σε περισσότερες από 200 χώρες.

5.5 Λογότυπο του μέσω κοινωνικής δικτύωσης LinkedIn.

Κεφάλαιο 6: Πρακτικό μέρος

6.1 Python

Η Python είναι διερμηνευόμενη (interprited), γενικού σκοπού (general-purpose) και υψηλού επιπέδου, γλώσσα προγραμματισμού. υποστηρίζει τόσο το διαδικαστικό (procedural programming) όσο και το αντικειμενοστραφές (object-oriented programming) προγραμματιστικό υπόδειγμα (programming paradigm). Είναι δυναμική γλώσσα προγραμματισμού (dynamically typed) και υποστηρίζει συλλογή απορριμμάτων (garbage collection ή GC).

Δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ (Guido van Rossum) στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) το 1989 και κυκλοφόρησε για πρώτη φορά το 1991.

Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. 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. Μειονεκτεί στο ότι επειδή είναι διερμηνευόμενη είναι πιο αργή από τις μεταγλωττιζόμενες (compiled) γλώσσες όπως η C και η C++. Γι’ αυτό το λόγο δεν είναι κατάλληλη για γραφή λειτουργικών συστημάτων.

Οι διερμηνευτές της Python είναι διαθέσιμοι για εγκατάσταση σε πολλά λειτουργικά συστήματα, επιτρέποντας στην Python την εκτέλεση κώδικα σε ευρεία γκάμα συστημάτων.

Οι λόγοι που επιλέξαμε τη python είναι 2. Ο 1^{ος} λόγος όπως αναφέραμε παραπάνω είναι πως η python είναι μια αρκετά απλή γλώσσα που προσφέρει πολλές δυνατότητες κυρίως ως προς τη διαχείριση αλφαριθμητικών (strings) και ο 2^{ος} αλλά κυριότερος λόγος είναι ότι υποστηρίζει μια πληθώρα βιβλιοθηκών. Μία από αυτές είναι το NLTK (Natural Language Toolkit) που είναι σχεδιασμένη ειδικά για εφαρμογές επεξεργασίας φυσικής γλώσσας και θα την αναλύσουμε παρακάτω.

6.1.1 IDLE

Για τη συγγραφή προγραμμάτων είναι απαραίτητος ένας κειμενογράφος ή ακόμα καλύτερα ένα ολοκληρωμένο περιβάλλον ανάπτυξης (Integrated Development Environment - IDE), το οποίο είναι ένα ειδικό λογισμικό για την ανάπτυξη εφαρμογών. Η Python έρχεται μαζί με ένα εύχρηστο και απλό περιβάλλον ανάπτυξης με την ονομασία IDLE. Το IDLE μας δίνει τη δυνατότητα να χρησιμοποιήσουμε διαδραστικά τον διερμηνευτή της γλώσσας, να γράψουμε και να επεξεργαστούμε προγράμματα, να τα αποθηκεύσουμε σε αρχεία, να τα εκτελέσουμε, να κάνουμε αποσφαλμάτωση.

6.2 NLTK

Το Toolkit για τη Φυσική Γλώσσα , ή πιο συχνά το NLTK , είναι μια σουίτα βιβλιοθηκών και προγραμμάτων για συμβολική και στατιστική επεξεργασία φυσικής γλώσσας (NLP) για αγγλικά, γραμμένη στη γλώσσα προγραμματισμού Python . Αναπτύχθηκε από τον Steven Bird και τον Edward Loper στο Τμήμα Πληροφορικής και Πληροφορικής του Πανεπιστημίου της Πενσυλβανίας . Το NLTK περιλαμβάνει γραφικές επιδείξεις και δείγματα δεδομένων. Προορίζεται να υποστηρίξει την έρευνα και τη διδασκαλία σε NLP ή σε στενά συνδεδεμένους τομείς, συμπεριλαμβανομένης της εμπειρικής γλωσσολογίας , της γνωστικής επιστήμης , της τεχνητής νοημοσύνης , της ανάκτησης πληροφοριών και της μηχανικής μάθησης . Παρέχει λεξιλογικούς πόρους, όπως το WordNet, μαζί με μια σειρά από βιβλιοθήκες επεξεργασίας κειμένου για ταξινόμηση, tokenization, stemming, tagging, parsing και semantic reasoning, wrappers for industrial-strength NLP βιβλιοθήκες και ένα ενεργό φόρουμ συζήτησης.

Το NLTK ενώ περιέχει ένα ευρύ σύνολο από λειτουργίες, παραμένει ένα εργαλείο που ασχολείται με το πεδίο της επεξεργασίας φυσικής γλώσσας και δεν μετατρέπεται σε σύστημα.

Παρακάτω παρουσιάζονται κάποιες από τις κύριες δυνατότητες που περιέχει το NLTK στην επεξεργασία φυσικής γλώσσας.

6.2.1 Tokenization

Με τη χρήση της βιβλιοθήκης nltk.tokenize μας δίνετε η δυνατότητα να μετατρέψουμε ένα κείμενο σε λίστα από tokens τόσο σε επίπεδο πρότασης όσο και σε επίπεδο λέξεων.

Για την εργασία μας χρησιμοποιήσαμε τη βιβλιοθήκη αυτή, εισάγοντάς την με τις εντολές:

>>> import nltk με αυτή την εντολή εισαγάγαμε το NLTK

>>> from nltk.tokenize import word_tokenize και με αυτήν εισαγάγαμε το tokenization

6.2.2 Part of Speech

Με τη βιβλιοθήκη αυτή δίνεται η δυνατότητα να αναγνωριστεί το μέρος του λόγου μιας πρότασης.

```
>>> text = word_tokenize("And now for something completely
different")
```

```
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something',
'NN'),
('completely', 'RB'), ('different', 'JJ')]
```

6.2.2.α Παράδειγμα κώδικα για αναγνώριση του μέρους του λόγου αφού έχουν εισαχθεί οι κατάλληλες βιβλιοθήκες.

POS TAG	DESCRIPTION
NNS	Ουσιαστικά σε πληθυντικό αριθμό.
NN	Ουσιαστικά σε ενικό αριθμό.
VBD	Ρήμα παρελθοντικό (π.χ. πήρε).
VBG	Ρήμα γερουνδιακό (τώρα).
VBN	Ρήμα παρελθοντικής συμμετοχής (κάποια στιγμή είχα κάνει κάτι)
VB2	Ρήμα για 3^{ov} άνθρωπο (π.χ. τραγουδάει, τώρα).
UH	Παρεμβολή (π.χ. errrrrrrmmmm).
WRB	Χρονική αντωνυμία (πότε, που).
PRP	Προσωπική αντωνυμία (εγώ, αυτός ,αυτή).
PRP$	Κτητική αντωνυμία (μου, δικά του).
RB	Επιρρήματα.
JJ	Επίθετο σε ενικό αριθμό.
FW	Ξένη λέξη.
NNS + NN	Όλα τα ουσιαστικά (ενικού και πληθυντικού αριθμού).
VBD + VBG + VBN + VB2	Όλων των ειδών ρημάτων.

6.2.2.β Πίνακας επεξήγησης χρησιμοποιούμενων σε αυτήν την εργασία Tag.

6.2.3 Stop Words

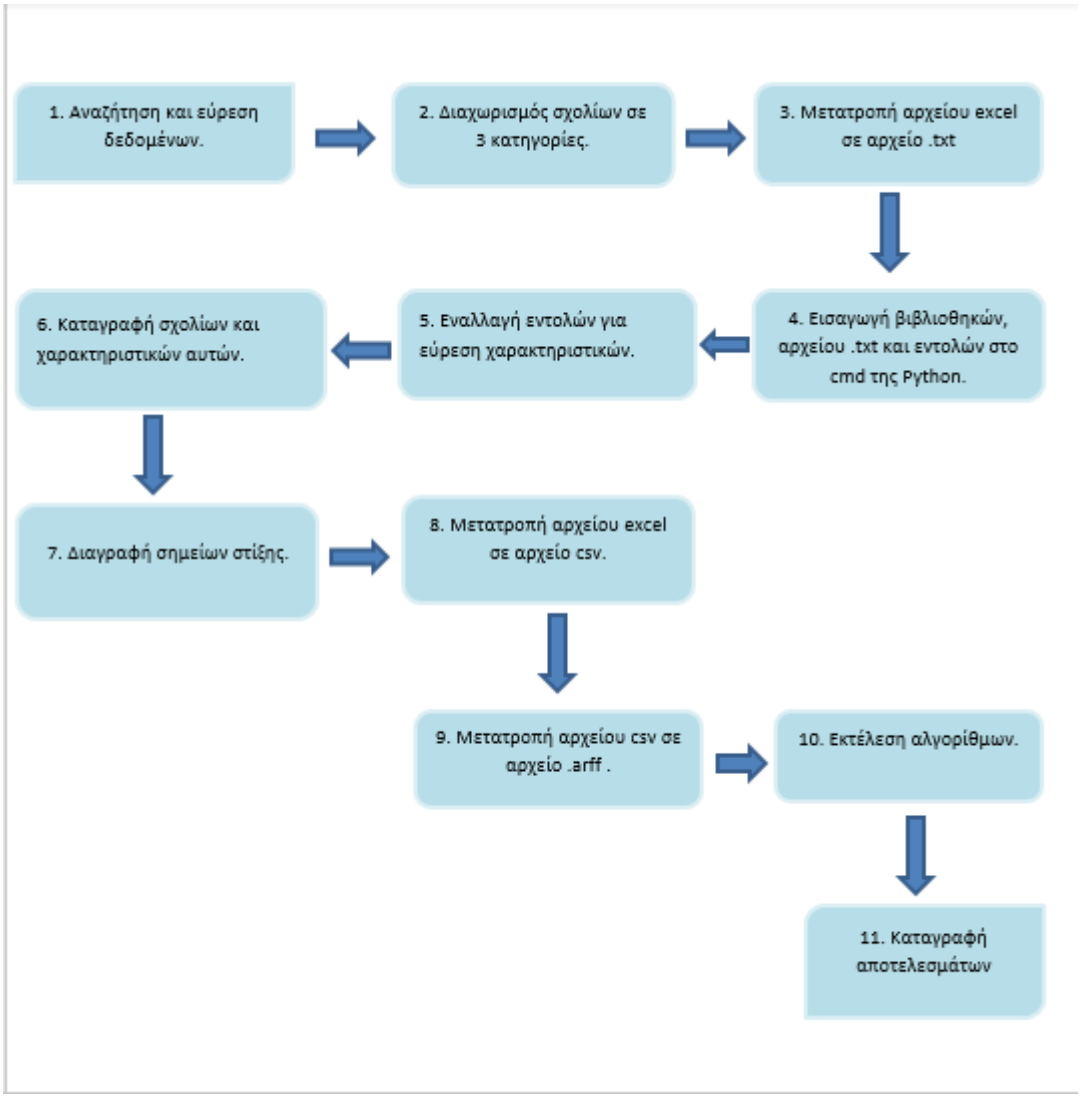
Οι Stop Words είναι ένα σύνολο από λέξεις, που δεν έχουν ιδιαίτερη βαρύτητα και το nltk δίνει τη δυνατότητα στο χρήστη να δώσει σαν είσοδο ένα αρχείο και να πάρει στην έξοδο το ίδιο αρχείο αλλά απαλλαγμένο από τις λέξεις αυτές.

6.3 Δεδομένα προς επεξεργασία

Σκοπός της εργασίας ήταν η δημιουργία ενός μοντέλου ανάλυσης της αγγλικής γλώσσας για σχόλια χρηστών, ώστε να ξεχωρίζει το συναίσθημα που είχε ο κάθε χρήστης όταν καταχωρούσε το σχόλιο του. Για αυτό χρησιμοποιήθηκαν έτοιμα δεδομένα. Συγκεκριμένα χρησιμοποιήθηκαν 1344 σχόλια χρηστών.

6.4 Σχεδιάγραμμα πτυχιακής εργασίας

Στο κεφάλαιο αυτό παρουσιάζεται ένα σχεδιάγραμμα με τα βήματα ένα προς ένα που ακολουθήσαμε για τη δημιουργία της πτυχιακή εργασίας μας.



6.4.1 Αναζήτηση και εύρεση δεδομένων

Για την έναρξη της πτυχιακής εργασίας μας χρειαζόμασταν δεδομένα και συγκεκριμένα χρειαζόμασταν σχόλια χρηστών από τα μέσα κοινωνικής δικτύωσης. Για την εύρεση των δεδομένων αυτών αναζητήσαμε στο internet και βρήκαμε 3 (τρία) έτοιμα αρχεία excel με περασμένα 1344 σχόλια χρηστών συνολικά στην αγγλική γλώσσα. Τα αρχεία ήταν 3 (τρία) γιατί τα δεδομένα ήταν χωρισμένα σε θετικά αρνητικά και ουδέτερα, σύμφωνα με το συναίσθημα που έβγαζε το κάθε σχόλιο.

6.4.2 Διαχωρισμός σχολίων σε 3 κατηγορίες

Αφού βρήκαμε τα δεδομένα που χρειαζόμασταν, ενσωματώσαμε τα τρία αρχεία σε ένα προσθέτοντας σε κάθε σχόλιο έναν αριθμό από το 1 (ένα) έως το 3 (τρία) ,που χαρακτήριζε την πολικότητα τους. Η πρόσθεση του αριθμού έγινε ως εξής: στα σχόλια που υπήρχαν στο αρχείο θετικών σχολίων βάλαμε τον αριθμό 3 (τρία), στα σχόλια που υπήρχαν στο αρχείο ουδέτερων βάλαμε τον αριθμό 2 (δύο) και στα σχόλια του αρχείου αρνητικών σχολίων βάλαμε τον αριθμό 1 (ένα). Με αυτό το τρόπο διαχωρίσαμε τα δεδομένα στις κατηγορίες 1 (ένα), 2 (δύο) και 3 (τρία) ανάλογα με τον αν ήταν θετικά, αρνητικά ή ουδέτερα συναισθηματικά φορτισμένα.

6.4.3 Μετατροπή αρχείου excel σε αρχείο .txt

Στη συνέχεια δημιουργήσαμε ένα αρχείο .txt το οποίο περιείχε μόνο τα σχόλια από το excel που είχαμε δημιουργήσει νωρίτερα. Αυτό το κάναμε για να μπορούμε να δίνουμε το αρχείο στη Python και να μας βγάζει σε ένα νέο αρχείο τον αριθμό των χαρακτηριστικών που επιθυμούσαμε για τη κάθε πρόταση. Αυτό η python το έκανε με συγκεκριμένες εντολές που θα παρουσιάσουμε και θα αναλύσουμε στο κεφάλαιο 6.4.4 .

6.4.4 Εισαγωγή βιβλιοθηκών, αρχείου .txt και εντολές στο cmd της Python

Για την εξαγωγή των χαρακτηριστικών εισαγάγαμε στη Python βιβλιοθήκες που χρειαζόμασταν με τις εξής εντολές:

```
from nltk.tag import pos_tag
from nltk.tokenize import word_tokenize
from collections import Counter
```

Ανοίξαμε το αρχείο .txt που είχαμε δημιουργήσει με την εντολή:

`rfile = open('sentinments_data.txt', 'r')`, δίνοντας το δικαίωμα read ('r') και δημιουργήσαμε ένα νέο αρχείο με δικαίωμα write ('w') με την εντολή:
`wfile = open('write.txt', 'w').`

Στη συνέχεια με ένα βρόχο επανάληψης για κάθε πρόταση που υπήρχε στο εισαγόμενο αρχείο η Python με τη βοήθεια των βιβλιοθηκών μετρούσε τα χαρακτηριστικά που εμείς θέλαμε.

```
for x in rfile:
  count = Counter([j for i,j in pos_tag(word_tokenize(x))])
  NN = count['FW']
  n = str(NN)
```

και τα εμφάνιζε στο νέο αρχείο που είχε δημιουργηθεί εκτελώντας τις εντολές:

```
  wfile.write('\n')
  wfile.write(n)
```

6.4.5 Εναλλαγή εντολών για εύρεση χαρακτηριστικών

Ανάλογα με ποια χαρακτηριστικά θέλαμε να μας εμφανίσει βάζαμε στο count το αντίστοιχο συμβολισμό τους που προέρχεται από το nltk και συγκεκριμένα από τη βιβλιοθήκη Part of Speech (pos). Τα σύμβολα που χρησιμοποιήσαμε αναλύονται παραπάνω στο κεφάλαιο 6.2.2 Για το συνδυασμό χαρακτηριστικών προσθέταμε μερικές εντολές ακόμη:

```
  PRP = count['PRP']
  WRB = count['WRB']
  PRP2 = count['PRP$']
  VBZ = count['VBZ']
  NN = PRP + PRP2 + WRB
```

Σύμφωνα με τις εντολές αυτές μετράει το κάθε χαρακτηριστικό μόνο του και μετά τα προσθέτει όλα μαζί εμφανίζοντας τελικά τον αριθμό που προέκυψε από τη πρόσθεση.

6.4.6 Καταγραφή σχολίων και χαρακτηριστικών αυτών

Βγάζοντας σαν αποτέλεσμα τον αριθμό των χαρακτηριστικών που επιθυμούσαμε κάθε φορά για κάθε πρόταση, παίρναμε το αποτέλεσμα αυτό και το αντιγράφαμε στο excel με αποτέλεσμα δίπλα από το κάθε σχόλιο να υπάρχει ο αριθμός κάθε χαρακτηριστικού που υπήρχε εντός αυτού του σχολίου.

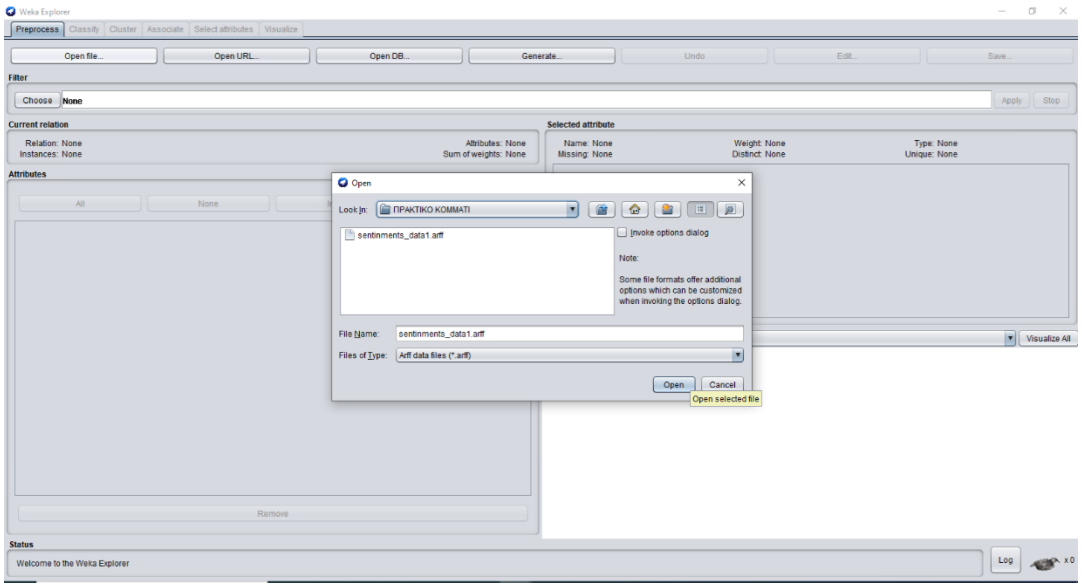
6.4.7 Διαγραφή σημείων στίξης και μετατροπές αρχείων

Αμέσως μετά διαγράψαμε τα σημεία στίξης από τα σχόλια για να μπορέσουμε να μετατρέψουμε το excel σε αρχείο . csv . Το αρχείο .csv δεν επιτρέπει τη χρήση σημείων στίξεων. Η μετατροπή αυτή ήταν αναπόφευκτη, αφού για να μπορέσουμε να δημιουργήσουμε αρχείο .arff που διαβάζει το Weka έπρεπε να έχουμε πρώτα το αρχείο .csv . Ύστερα έχοντας το αρχείο .csv μετατρέψαμε και αυτό σε αρχείο .arff .

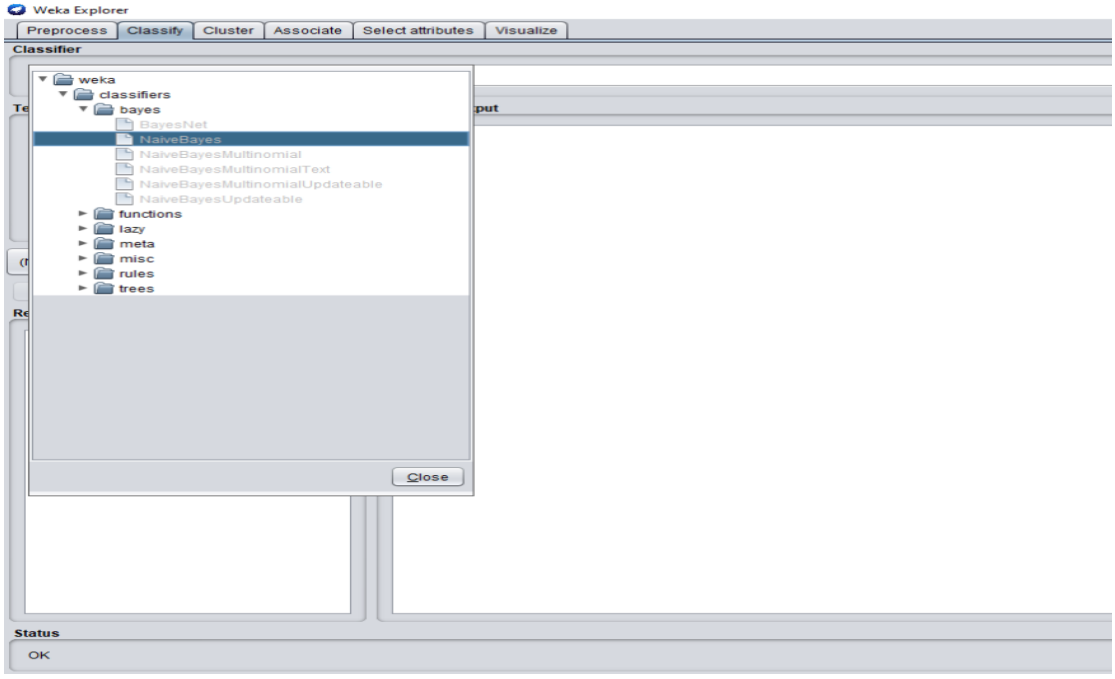
Οι μετατροπές αυτές έγιναν ως εξής: Από excel σε .csv το μετατρέψαμε κάνοντας χρήση ενός προγράμματος που βρήκαμε στο internet, ενώ από .csv σε .arff το κάναμε μέσω του Weka που μας παρείχε αυτή τη δυνατότητα.

6.4.8 Εκτέλεση αλγορίθμων

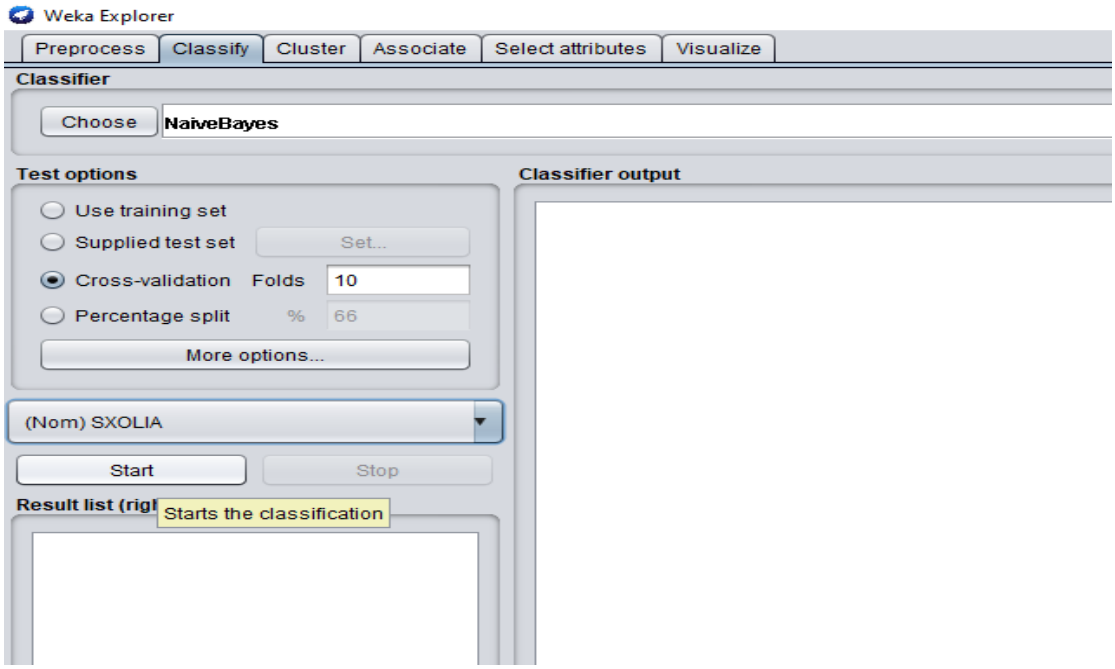
Για να γίνει η εκτέλεση κάθε αλγορίθμου αρχικά ανοίγαμε το αρχείο .arff εντός του Weka, όπως φαίνεται στην εικόνα:



Έπειτα, στη καρτέλα classify για κάθε αλγόριθμο επιλέγαμε αυτόν που θέλαμε να εκτελεστεί:



Και τέλος πατούσαμε “Start” για έναρξη εκτέλεσης:

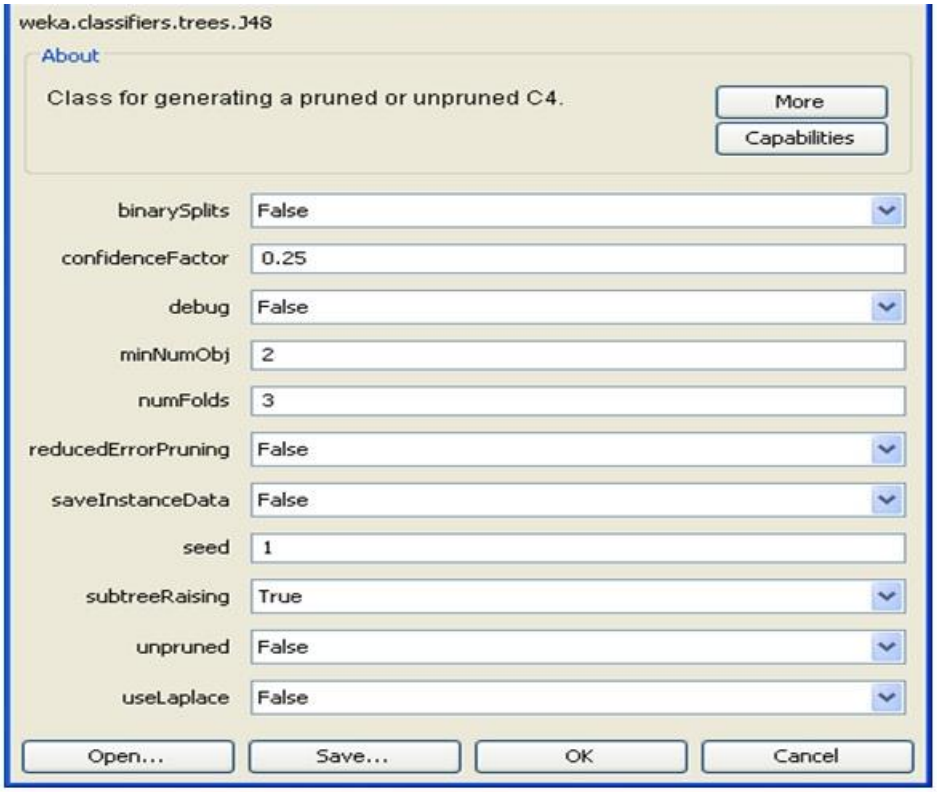


6.4.9 Καταγραφή αποτελεσμάτων

Σαν τελευταίο βήμα του πρακτικού μέρους της πτυχιακής μας ήταν η καταγραφή των αποτελεσμάτων για κάθε αλγόριθμο που εκτελέσαμε, προκειμένου να γίνει η σύγκριση (κεφάλαιο 6.8).

6.5 Ο αλγόριθμος J48

Ο αλγόριθμος J48 αναπτύχθηκε το 1993 και στην ουσία είναι μια εξέλιξη του C4.5 . Πως ακριβώς δουλεύει όμως; Ο αλγόριθμος αυτός παράγει ένα δέντρο απόφασης και διαχειρίζεται κατηγορηματικές και συνεχείς τιμές. Εφαρμόζεται από πάνω προς τα κάτω με επαναληπτική διαδικασία υλοποίησης της μεθόδου «διαίρει και βασίλευε». Χρησιμοποιείται για να εντοπίσει το ολικό βέλτιστο, αναζητώντας τοπικά βέλτιστες επιλογές. Στη συνέχεια παρουσιάζεται μία εικόνα, η οποία δείχνει τις παραμέτρους που χρησιμοποιεί κατά την πειραματική εφαρμογή ο εξεταζόμενος αλγόριθμος.



6.5.α Πίνακας παραμέτρων αλγορίθμου J48

Η παράμετρος *Binary Splits* χρησιμοποιείται εάν θέλουμε να χρησιμοποιηθεί δυαδικός διαμερισμός στα ονομαστικά χαρακτηριστικά όταν κατασκευάζονται δένδρα.

Η *Confidence Factor* παράμετρος αποτελεί τον παράγοντα εμπιστοσύνης, ο οποίος χρησιμοποιείται για «κλάδεμα» (μικρότερες τιμές επιφέρει περισσότερο κλάδεμα).

Η *Min Num Obj* δείχνει τον ελάχιστο αριθμό παραδειγμάτων σε κάθε φύλλο,ενώ η παράμετρος *Num Folds* προσδιορίζει το ποσό των δεδομένων που χρησιμοποιούνται για την ελάττωση του σφάλματος «κλαδέματος». Η μια επανάληψη χρησιμοποιείται για «κλάδεμα», οι υπόλοιπες για την ανάπτυξη του δένδρου.

Η *Reduced Error Pruning* επιλέγεται όταν θέλουμε να χρησιμοποιηθεί η ελάττωση σφάλματος «κλαδέματος» αντί για το C4.5 κλαδεμένο δένδρο.

Η παράμετρος *Save Instance Data* επιλέγεται όταν θέλουμε να αποθηκεύσουμε τα δεδομένα εκπαίδευσης για απεικόνιση.

Στην παράμετρο *Seed,* ο σπόρος χρησιμοποιείται για την τυχαιοποίηση των δεδομένων όταν χρησιμοποιείται η ελάττωση σφάλματος «κλαδέματος».

Η παράμετρος *Subtree Raising* επιλέγεται εάν θέλουμε να λαμβάνεται υπ' όψιν η λειτουργία της ανάπτυξης του υπό-δένδρου κατά το «κλάδεμα».

Η *Unpruned* επιλέγεται εάν θέλουμε να εκτελεστεί το «κλάδεμα» και

τέλος, η παράμετρος *Use Laplace* χρησιμοποιείται όταν θέλουμε τα μέτρα των φύλλων να είναι ομαλοποιημένα με βάση του κριτηρίου *Laplace*.

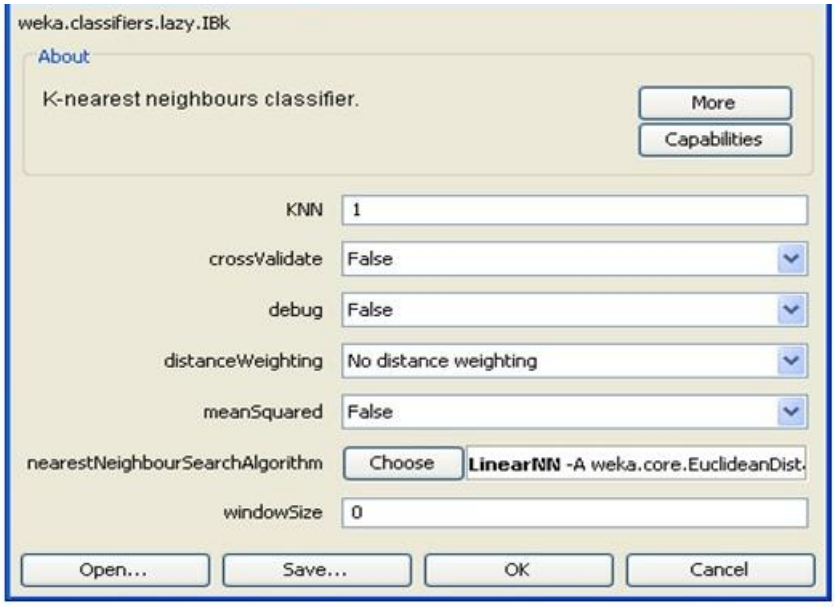
6.6 Ο αλγόριθμος Ibk

Ο IBK είναι από τους πιο σημαντικούς αλγόριθμους της κατηγορίας της ταξινόμησης. Κρατά μία πλήρη μνήμη των στιγμιότυπων κατάρτισης και ταξινομεί τις νέες περιπτώσεις χρησιμοποιώντας τις περισσότερο παρόμοιες περιπτώσεις κατάρτισης. Η νέα περίπτωση ταξινομείται μετά από την εύρεση της περίπτωσης με την μεγαλύτερη ομοιότητα όπου της προσαρτάται η αντίστοιχη κλάση. Το μειονέκτημα του αλγόριθμου αυτού είναι ότι διατηρεί μια μεγάλη μνήμη για την αποθήκευση των στιγμιότυπων κατάρτισης. Για να προβλέψει την κλάση του καινούργιου στιγμιότυπου ο IBK χρησιμοποιεί την συνάρτηση ομοιότητας που ορίζεται ως:

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)}$$

Ο αλγόριθμος αυτός (αλγόριθμος k-πλησιέστερων γειτόνων (k-NN)) χρησιμοποιείται ως μια μη παραμετρική* μέθοδος για την ταξινόμηση και την παλινδρόμηση*. Στις περισσότερες περιπτώσεις η είσοδος αποτελείται από τα k πλησιέστερα παραδείγματα κατάρτισης και η έξοδος εξαρτάται από τον αν το k-NN* χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση.

Ο αριθμός των παραδειγμάτων εκπαίδευσης του ταξινομητή μπορούν να περιοριστούν από την επιλογή του παραθύρου παραμέτρων που αναλύονται παρακάτω ,ενώ όταν προθέτονται νέα παραδείγματα εκπαίδευσης, τα παλιά απομακρύνονται ώστε να διατηρηθεί ο αριθμός των παραδειγμάτων εκπαίδευσης.



6.6.α Πίνακας παραμέτρων αλγορίθμου Ibk

Η παράμετρος *Entropic Auto Blend* επιλέγεται εάν θέλουμε να χρησιμοποιηθεί η εντροπία βασισμένη στο δείγμα, ενώ η παράμετρος *Global Blend* αναφέρεται στο σφαιρικό μίγμα και οι τιμές της κυμαίνονται από το 0 έως το 100. Επίσης με την παράμετρο *Missing Mode* προσδιορίζεται πώς οι ελλιπείς τιμές των χαρακτηριστικών μεταχειρίζονται. Έχουμε τέσσερις επιλογές:

α) αγνοούμε τα παραδείγματα με ελλιπείς τιμές (ignore the instances with missing values),

β) χειριζόμαστε τις ελλιπείς τιμές ως μέγιστες τιμές (treat missing values as maximally different),

γ) εξομαλύνουμε τα χαρακτηριστικά (normalize over the attributes) και

δ) υπολογίζουμε τον μέσο όρο των καμπυλών εντροπίας (average column entropy curves) την οποία χρησιμοποιούμε στην περίπτωσή μας.

*Μη παραμετρική = Δεν βασίζονται σε υποθέσεις ότι τα δεδομένα προέρχονται από μια δεδομένη παραμετρική οικογένεια κατανομών πιθανοτήτων

*Η ανάλυση παλινδρόμησης είναι ένα σύνολο στατιστικών διεργασιών για την εκτίμηση των σχέσεων μεταξύ μιας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών.

*k-NN είναι ένας τύπος εκμάθησης βασισμένης σε στιγμιότυπα .

A	B	C	<--Classified as

6.7 Ο αλγόριθμος Decision Stump

Ο αλγόριθμος Decision Stump είναι ένα μοντέλο μηχανικής μάθησης που αποτελείται από ένα δέντρο απόφασης ενός επιπέδου. Δηλαδή, είναι ένα δέντρο απόφασης με έναν εσωτερικό κόμβο (τη ρίζα) που συνδέεται άμεσα με τους τερματικούς κόμβους (τα φύλλα του).Ο Decision Stump κάνει μια πρόβλεψη με βάση την αξία ενός μόνο χαρακτηριστικού εισόδου. Μερικές φορές ονομάζονται επίσης 1-rules.

Ανάλογα με τον τύπο της δυνατότητας εισαγωγής, είναι δυνατές αρκετές παραλλαγές. Για ονομαστικά χαρακτηριστικά, μπορεί να κατασκευαστεί ένας κολοβός που περιέχει ένα φύλλο για κάθε πιθανή τιμή χαρακτηριστικού ή ένα κορμό με τα δύο φύλλα, ένα από τα οποία αντιστοιχεί σε κάποια επιλεγμένη κατηγορία και το άλλο φύλλο σε όλες τις άλλες κατηγορίες .Για τα δυαδικά χαρακτηριστικά αυτά τα δύο σχήματα είναι πανομοιότυπα. Μια τιμή που λείπει μπορεί να θεωρηθεί ως μια άλλη κατηγορία. Για συνεχή χαρακτηριστικά, συνήθως, έχει επιλεγεί κάποια τιμή χαρακτηριστικού κατωφλίου και ο κολοβός περιέχει δύο φύλλα - για τιμές κάτω και πάνω από το όριο. Ωστόσο, σπάνια μπορούν να επιλεγούν πολλαπλά όρια και ο κορμός να περιέχει τρία ή περισσότερα φύλλα.

Τα κουμπιά απόφασης συχνά χρησιμοποιούνται ως συνιστώσες, που ονομάζονται «αδύναμοι μαθητές» ή «βασικοί μαθητές» , σε τεχνικές συλλογικής μάθησης, όπως η σάρωση και η ενίσχυση.

6.8 Αποτελέσματα και σύγκριση βασικών αλγορίθμων

Βάζοντας τα σχόλια χρηστών, τα χαρακτηριστικά του κάθε σχολίου καθώς και την κατηγορία στην οποία ανήκει κάθε σχόλιο (δηλ. το αρχείο .arff) ως είσοδο στο λογισμικό Weka, επιλέγοντας τους αντίστοιχους αλγορίθμους ταξινόμησης προκύπτει ότι ο αλγόριθμος J48 είναι ένας από τους αλγόριθμους με το μεγαλύτερο ποσοστό επιτυχίας (56,06%) στην εργασία μας, καθώς επίσης και ο Ibk με λίγο μικρότερο ποσοστό επιτυχίας (52,1%), ενώ ο Decision Stump είναι αλγόριθμος με το χαμηλότερο ποσοστό επιτυχίας(49,73).

6.8.1 Αποτελέσματα αλγορίθμου J48

Στο σύνολο των 1343 σχολίων που χρησιμοποιήσαμε στην συγκεκριμένη εργασία ο αλγόριθμος J48 κατηγοριοποίησε σωστά 753 σχόλια σε 3 κατηγορίες πετυχαίνοντας ποσοστό επιτυχίας 56,06% και αντίστοιχα ποσοστό αποτυχίας 43,94%. Παρακάτω φαίνεται η τελική κατηγοριοποίηση των σχολίων που έκανε ο αλγόριθμος.

240	39	109	A = Class 1
125	261	82	B = Class 2
95	140	252	C = Class 3

Πίνακας 6.8.1 Κατηγοριοποίηση σχολίων ανά κλάση του J48

Στον πίνακα 6.8.1 δείχνει ότι η κλάση 1 έχει σύνολο 388 σχόλια εκ των οποίων τα 240 κατηγοριοποιήθηκαν σωστά στην κλάση 1 ενώ τα υπόλοιπα 148 κατηγοριοποιήθηκαν εσφαλμένα στις άλλες κλάσεις. Το ποσοστό ακριβείας για τη κλάση 1 ήταν 61,85%. Αντίστοιχα για την κλάση 2 κατηγοριοποιήθηκαν σωστά 261 σχόλια από τα 468 που είχε ως σύνολο και τα υπόλοιπα κατηγοριοποιήθηκαν εσφαλμένα. Το ποσοστό ακρίβειας ήταν 55,77% . Για την κλάση 3 δόθηκαν συνολικά 487 σχόλια χρηστών και κατηγοριοποιήθηκαν σωστά τα 252 ενώ τα υπόλοιπα 235 λάθος. Το ποσοστό ακρίβειας της κλάσης αυτής ήταν 51,74%.

6.8.2 Αποτελέσματα αλγορίθμου Ibk

Στο σύνολο των 1343 σχολίων ο αλγόριθμος Ibk κατηγοριοποίησε σωστά 700 σχόλια σε 3 κατηγορίες πετυχαίνοντας ποσοστό επιτυχίας 52,1% και αντίστοιχα ποσοστό αποτυχίας 47,9%. Παρακάτω φαίνεται η τελική κατηγοριοποίηση των σχολίων που έκανε ο αλγόριθμος.

A	**B**	**C**	**<--Classified as**
215	145	42	A = Class 1
85	275	166	B = Class 2
93	112	210	C = Class 3

Πίνακας 6.8.2 Κατηγοριοποίηση σχολίων ανά κλάση του Ibk

Στον πίνακα 6.8.2 δείχνει ότι η κλάση 1 έχει σύνολο 402 σχόλια εκ των οποίων τα 215 κατηγοριοποιήθηκαν σωστά στην κλάση 1 ενώ τα υπόλοιπα 287 κατηγοριοποιήθηκαν εσφαλμένα στις άλλες κλάσεις. Το ποσοστό ακριβείας για τη κλάση 1 ήταν 53,48%. Αντίστοιχα για την κλάση 2 κατηγοριοποιήθηκαν σωστά 85 σχόλια από τα 526 που είχε ως σύνολο και τα υπόλοιπα κατηγοριοποιήθηκαν εσφαλμένα. Το ποσοστό ακρίβειας ήταν 52,28% . Για την κλάση 3 δόθηκαν συνολικά 415 σχόλια χρηστών και κατηγοριοποιήθηκαν σωστά τα 210 ενώ τα υπόλοιπα 205 λάθος. Το ποσοστό ακρίβειας της κλάσης αυτής ήταν 50,60%.

6.8.3 Αποτελέσματα αλγορίθμου Decision Stump

Στο σύνολο των 1343 σχολίων ο αλγόριθμος Decision Stump κατηγοριοποίησε σωστά 668 σχόλια στις ίδιες (3) κατηγορίες πετυχαίνοντας ποσοστό επιτυχίας 49,73% και αντίστοιχα ποσοστό αποτυχίας 50,26%. Παρακάτω φαίνεται η τελική κατηγοριοποίηση των σχολίων που έκανε ο αλγόριθμος.

A	B	C	<--Classified as
200	136	0	A = Class 1
68	268	0	B = Class 2
0	136	200	C = Class 3

Πίνακας 6.8.3 Κατηγοριοποίηση σχολίων ανά κλάση του Decision Stump

Στον πίνακα 6.8.3 δείχνει ότι η κλάση 1 έχει σύνολο 336 σχόλια εκ των οποίων τα 200 κατηγοριοποιήθηκαν σωστά στην κλάση 1 ενώ τα υπόλοιπα 136 κατηγοριοποιήθηκαν εσφαλμένα στη κλάση 2. Το ποσοστό ακριβείας για τη κλάση 1 ήταν 59,52%. Αντίστοιχα για την κλάση 2 κατηγοριοποιήθηκαν σωστά 268 σχόλια από τα 336 που είχε ως σύνολο και τα υπόλοιπα κατηγοριοποιήθηκαν εσφαλμένα στη κλάση 1.Το ποσοστό ακρίβειας ήταν 79,76% . Ωστόσο και για την κλάση 3 δόθηκαν συνολικά 336 σχόλια χρηστών και κατηγοριοποιήθηκαν σωστά τα 200 ενώ τα υπόλοιπα 136 λάθος. Το ποσοστό ακρίβειας της κλάσης αυτής ήταν 59,52%.

Συνοψίζοντας, στη παρούσα εργασία δημιουργήθηκε ένα μοντέλο ανάλυσης συναισθήματος στην αγγλική γλώσσα βασισμένο στα χαρακτηριστικά κάθε σχολίου και την κατηγοριοποίηση σε θετικό, αρνητικό ή ουδέτερο.

Βιβλιογραφία

1. https://blog.cambridgespark.com/50-free-machine-learning-datasets-sentiment-analysis-b9388f79c124
2. https://www.softaware.gr/%CF%85%CF%80%CE%B7%CF%81%CE%B5%CF%83%CE%B9%CE%B5%CF%82-%CF%80%CE%BB%CE%B7%CF%81%CE%BF%CF%86%CE%BF%CF%81%CE%B9%CE%BA%CE%B7%CF%82/%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%B7-%CE%BC%CE%B1%CE%B8%CE%B7%CF%83%CE%B7/https://el.wikipedia.org/wiki/Weka_(%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7)
3. https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf
4. https://el.wikipedia.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7#%CE%98%CE%B5%CF%89%CF%81%CE%AF%CE%B1
5. http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/13134/1/DT2016-0114.pdf
6. http://ir.lib.uth.gr/bitstream/handle/11615/41387/10471.pdf?sequence=1&isAllowed=y
7. http://okeanis.lib2.uniwa.gr/xmlui/bitstream/handle/123456789/3779/%CE%A0%CE%A4%CE%A5%CE%A7%CE%99%CE%91%CE%9A%CE%97_%CE%95%CE%A1%CE%93%CE%91%CE%A3%CE%99%CE%91.pdf?sequence=1&isAllowed=y
8. http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/9421/Dimoulis_Nikolaos.pdf?sequence=1&isAllowed=y
9. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
10. https://en.wikipedia.org/wiki/Regression_analysis
11. https://en.wikipedia.org/wiki/Pattern_recognition
12. https://en.wikipedia.org/wiki/Nonparametric_statistics
13. https://nemertes.lis.upatras.gr/jspui/bitstream/10889/8561/1/datamining%20medical%20data.pdf
14. https://www.csc.com.gr/machine-learning-%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE-%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7-%CF%84%CE%B9-%CE%B5%CE%AF%CE%BD%CE%B1%CE%B9/
15. https://www.insider.gr/epiheiriseis/tehnologia/92698/10-kathimerines-efarmoges-tis-vathias-mathisis
16. https://docplayer.gr/55226885-Analysi-synaisthematos-se-koinonika-iktya-shetika-me-ta-oikonomika-metra-stin-ellada.html

17. https://el.wikipedia.org/wiki/Facebook
18. https://el.wikipedia.org/wiki/Twitter#cite_note-4
19. https://el.wikipedia.org/wiki/Instagram
20. https://el.wikipedia.org/wiki/LinkedIn
21. https://el.wikipedia.org/wiki/Python
22. https://en.wikipedia.org/wiki/Natural_Language_Toolkit
23. http://artemis.library.tuc.gr/DT2009-0140/DT2009-0140.pdf
24. https://en.wikipedia.org/wiki/Decision_stump