ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ (ΤΕΙ) ΑΘΗΝΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ

Τεχνικές αυτόματης κατηγοριοποίησης στις ψηφιακές βιβλιοθήκες

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ευφροσύνη Ε. Βοργιά

Δρ Ιωάννης Τριανταφύλλου

Αθήνα, Νοέμβριος 2016

Ευχαριστίες

Ο άνθρωπος που είμαι σήμερα είναι αποτέλεσμα των ιστοριών που έχω δει, ακούσει και διαβάσει, των καταστάσεων που έχω βιώσει και των ανθρώπων που έχω γνωρίσει. Είμαι χαρούμενη για όσα έχω ζήσει και ευγνώμων σε όσους βρέθηκαν στο δρόμο μου. Δυστυχώς είναι αδύνατο να τους αναφέρω όλους, αλλά μπορώ να ευχαριστήσω τους βασικούς.

Πρώτα θα ήθελα να ευχαριστήσω τον επιβλέποντά μου Δρ Ιωάννη Τριανταφύλλου. Το θέμα της παρούσας πτυχιακής ήταν δική του ιδέα και βασίστηκε σε προηγούμενη δουλειά του. Έδειξε μεγάλη υπομονή για την ολοκλήρωσή της και με βοήθησε.

Έπειτα, ευχαριστώ τις καλύτερές μου φίλες Ελένη Μπάσουλα, Μαρία Βάκου, Αικατερίνη Βινιεράτου και τις οικογένειές τους. Από την πρώτη στιγμή που με γνώρισαν με στήριξαν σαν δικό τους άνθρωπο και ήταν ανεκτικές κάθε φορά που τους έλεγα «Δεν μπορώ να βγω, έχω διάβασμα».

Τέλος, ευχαριστώ την οικογένειά μου, τον πατέρα μου Λευτέρη που μας άφησε νωρίς, τα αδέλφια μου, τα ανίψια μου και κυρίως τη μαμά μου Όλγα. Δίχως τη δική της αγάπη είναι ο κόσμος πιο μικρός!

Περιεχόμενα

Περίληψη

Η υψηλή ταχύτητα και ο όγκος εισαγωγής τεκμηρίων στις ψηφιακές βιβλιοθήκες απαιτούν την άμεση τεκμηρίωση του υλικού. Αν και σχεδόν όλες οι εργασίες είναι πλέον αυτοματοποιημένες, η ταξινόμηση γίνεται ακόμα με την ανθρώπινη εμπλοκή. Η παρούσα πτυχιακή εργασία προτείνει την αυτόματη κατηγοριοποίηση των τεκμηρίων με βάση τις λέξεις των περιλήψεών τους, ενώ μελετά τεχνικές κατηγοριοποίησης (classification) και συσταδοποίησης (clustering) για την επίτευξη της εργασίας.

Η έρευνα στοχεύει στην εισαγωγή μιας νέας μετρικής βαρύτητας, της DEVMAX.DF, η οποία συγκρίνεται με την ήδη υπάρχουσα μετρική TF.IDF. Οι μετρικές συνδυάζονται με τη χρήση της δυαδικής εμφάνισης και της πραγματικής συχνότητας των λέξεων. Η αποτελεσματικότητά τους εξετάζεται πάνω σε διάφορους αλγορίθμους κατηγοριοποίησης και συσταδοποίησης.

Τα δεδομένα λήφθηκαν από 718 περιλήψεις επιστημονικών μελετών από 9 ακαδημαϊκές ψηφιακές βιβλιοθήκες. Μετά από την επεξεργασία τους, εισήχθησαν στο πρόγραμμα WEKA για την εξαγωγή των αποτελεσμάτων.

Η κατηγοριοποίηση απέφερε ανώτατο F-score ~97%, ενώ η συσταδοποίηση λανθασμένων παραδειγμάτων έφτασε ~4,50%. Εκ των δύο μετρικών, η DEVMAX.DF απέδωσε καλύτερα από την TF.IDF, ενώ η δυαδική εμφάνιση των λέξεων ενίσχυσε περισσότερο τα αποτελέσματα από ότι η εμφάνιση συχνότητας.

Λέξεις κλειδιά

Ψηφιακές Βιβλιοθήκες, Κατηγοριοποίηση, Συσταδοποίηση, WEKA, Μετρικές Βαρύτητας

Abstract

The high speed and the import volume of documents in digital libraries require immediate documentation. Although almost all the tasks are automated, the classification is still conducted by humans. This thesis proposes the automated categorization of documents based on the words of abstracts, while studies classification and clustering techniques to achieve it.

The research aims to introduce a new weighting metric, DEVMAX.DF, which is compared to the already existing metric TF.IDF. The metrics are combined with the use of the binary word display and the actual term frequency. Their efficiency is tested on different classification and clustering algorithms.

The data were gathered from 718 abstracts of scientific studies from 9 academic digital libraries. After processing, the data were imported to WEKA for the final results.

Classification yielded an F-score ~ 97%, while clustering reached ~4.50% of incorrectly clustered instances. Of the two metrics, DEVMAX.DF performed better than TF.IDF, while the binary display strengthened the results more than the term frequency.

Keywords

Digital Libraries, Classification, Clustering, WEKA, Weighting methods

1. Εισαγωγή

Ζούμε μια επανάσταση, την ψηφιακή επανάσταση που ξεκίνησε τη δεκαετία του 1980 και συνεχίζεται με ραγδαίους ρυθμούς σήμερα. Το ψηφιακό περιεχόμενο έχει μπει στην καθημερινότητα κάθε ανθρώπου αντικαθιστώντας το αναλογικό. Στην κοινωνία που παράγει όλο και περισσότερα ψηφιακά αντικείμενα, ή ψηφιοποιεί τα αναλογικά, οι ψηφιακές βιβλιοθήκες είναι απαραίτητες για την εύρεση ποιοτικού, έγκυρου και σύγχρονου υλικού.

Ο όγκος των παραγόμενων αντικειμένων και ο ρυθμός εισαγωγής τους απαιτούν την άμεση τεκμηρίωσή του. Αν και η αυτοματοποίηση βοηθά στην ταχύτερη περάτωση αυτής της εργασίας, δεν έχει γίνει πλήρως εκμεταλλεύσιμη. Ένα κομμάτι της τεκμηρίωσης είναι και η ταξινόμηση, η οποία ακόμα γίνεται με την ανθρώπινη εμπλοκή και τη χρήση συστημάτων ταξινόμησης όπως το Dewey Decimal Classification ή το Library of Congress Classification System. Όμως, υπό αυτές τις συνθήκες η ταξινόμηση απαιτεί πολλούς ανθρώπους και πολλές ώρες.

Η παρούσα πτυχιακή προτείνει και μελετά την αποδοτικότητα τεχνικών αυτόματης κατηγοριοποίησης βασιζόμενες στην Μηχανική Μάθηση για την αντιμετώπιση του προβλήματος. Πρόκειται για τις διαδικασίες της κατηγοριοποίησης (classification) και της συσταδοποίησης (clustering).

Οι στόχοι της έρευνας είναι:

• Η εισαγωγή μιας νέας μετρικής βαρύτητας, της DEVMAX.DF. Η συγκεκριμένη μετρική δημιουργείται και προτείνεται για την επιλογή εύστοχων λέξεων, βελτιώνοντας τα αποτελέσματα.

• Η σύγκριση της DEVMAX.DF με την ήδη υπάρχουσα μετρική TF.IDF. Οι μετρικές συνδυάζονται με τη χρήση της δυαδικής εμφάνισης και της πραγματικής συχνότητας εμφάνισης των λέξεων.

• Η εφαρμογή των αλγορίθμων κατηγοριοποίησης (classification) και συσταδοποίησης (clustering) και η σύγκριση της αποδοτικότητάς τους.

Οι περιλήψεις των επιστημονικών μελετών αποτυπώνουν τα κομβικά σημεία με την κατάλληλη ορολογία καταλαμβάνοντας λιγότερο όγκο σχετικά με το πλήρες κείμενο. Λαμβάνοντας υπόψη αυτή την ιδιότητα, γίνεται συγκέντρωση και επεξεργασία των περιλήψεων. Σκοπός είναι η απομόνωση εύστοχων λέξεων που θα χρησιμοποιηθούν για την εξέταση των τεχνικών κατηγοριοποίησης και συσταδοποίησης μέσω του προγράμματος WEKA.

Αρχικά, συλλέγονται 718 περιλήψεις από 9 ελληνικές ακαδημαϊκές ψηφιακές βιβλιοθήκες. Έπειτα, κατά το χειρισμό του συνόλου των κειμένων, αφαιρούνται οι τετριμμένες/λειτουργικές λέξεις (stop words), ενώ οι εναπομένουσες περιορίζονται στο κυρίως θέμα τους και συντάσσονται σε ομάδες βάσει αυτού υπό έναν όρο-αντιπρόσωπο.

Στη συνέχεια, εφαρμόζονται οι μετρικές βαρύτητας TF.IDF και DEVMAX.DF σε συνδυασμό με τη δυαδική εμφάνιση των λέξεων, όπου δείχνει την ύπαρξη ή την απουσία μιας λέξης στην περίληψη, και την πραγματική συχνότητα εμφάνισης των λέξεων. Στο τελικό στάδιο γίνεται χρήση του προγράμματος WEKA για την εφαρμογή των τεχνικών classification και clustering και την εξαγωγή των αποτελεσμάτων.

Τέλος, ένα μέρος της μελέτης θα δημοσιευθεί συντόμως (Vorgia et al., 2017). Η ιδέα όλης της παραπάνω έρευνας βασίστηκε σε προηγούμενη δουλειά των Triantafyllou et al. (2014).

2. Βιβλιογραφική ανασκόπηση

2.1. Ψηφιακές βιβλιοθήκες και αποθετήρια

Η ψηφιακή βιβλιοθήκη ορίζεται ως μια «ηλεκτρονική συλλογή ψηφιακών αντικειμένων, εξασφαλισμένης ποιότητας, τα οποία δημιουργούνται ή συλλέγονται και διαχειρίζονται σύμφωνα με παγκοσμίως αποδεκτές αρχές για την ανάπτυξη της συλλογής ενώ είναι προσβάσιμα με ένα λογικό και βιώσιμο τρόπο, υποστηριζόμενα από υπηρεσίες απαραίτητες για να επιτρέψουν στους χρήστες να ανακτήσουν και να εκμεταλλευτούν τους πόρους» (IFLA & UNESCO, 2014, σ. 1).

Στην ίδια λογική με τις ψηφιακές βιβλιοθήκες είναι και τα αποθετήρια. Επιτρέπουν στους «ερευνητές, συγγραφείς και τα μέλη των ακαδημαϊκών ιδρυμάτων να καταθέτουν την συγγραφική τους παραγωγή για φύλαξη, κρίση, τεκμηρίωση και παραγωγή μεταδεδομένων, διάχυση στο διαδίκτυο και χρήση» (Κυριάκη - Μάνεση & Κουλούρης, 2015, σ. 11). Στην κατηγορία των αποθετηρίων μπορούν να συμπεριληφθούν ακόμα και βάσεις όπου διατηρούνται αρχεία ή και δεδομένα.

Το 1991 αποτελεί ημερομηνία ορόσημο για τις ψηφιακές βιβλιοθήκες, καθώς τότε ξεκίνησε η λειτουργία του E-prints archive ή, όπως ονομάζεται σήμερα, arXiv στο Los Amos Laboratory (Ginsparg, 1994; Candela et al., 2011; Καπιδάκης, 2014). Το arXiv είναι η πρώτη ψηφιακή βιβλιοθήκη, σύμφωνα με τη σημερινή μορφή τους, ενώ εξυπηρετεί τις ανάγκες των ερευνητών στις θετικές και τεχνολογικές επιστήμες, όπως τη Φυσική, τα Μαθηματικά, την Πληροφορική, τη Βιολογία, τα Οικονομικά και τη Στατιστική, παρέχοντας ανοικτή πρόσβαση σε περισσότερα από 1.000.000 άρθρα. Υπεύθυνο για τη λειτουργία του είναι το Πανεπιστήμιο Cornell, ενώ συγχρηματοδοτείται από τη βιβλιοθήκη του πανεπιστημίου, το ίδρυμα Simons και άλλους φορείς (arXiv, 2016).

Το arXiv είναι μία εκ των πολλών ψηφιακών βιβλιοθηκών που υπάρχουν παγκοσμίως. Άλλα μεγάλα συνεργατικά σχέδια παρόμοιου βεληνεκούς είναι το

HathiTrust[1], το Internet Archive[2], η Europeana[3], το PANDORA Archive[4] και η World Digital Library[5].

Ωστόσο, πρέπει να ληφθεί υπόψη η ύπαρξη διαφόρων άλλων μικρότερων συλλογών από:

- Χιλιάδες πανεπιστημιακές βιβλιοθήκες και επιστημονικά κέντρα για την κατάθεση διατριβών, άρθρων και ακατέργαστων δεδομένων (raw data), όπως το αποθετήριο του Πανεπιστήμιο Πειραιά[6].
- Πολλά μουσεία και αρχειακές υπηρεσίες που ψηφιοποιούν και αναρτούν μέρη των συλλογών τους, όπως το Morgan Museum and Library[7].
- Κυβερνητικούς και άλλους οργανισμούς οι οποίοι παρέχουν ελεύθερα τα τεκμήριά τους, όπως το αποθετήριο του Ελεύθερου Λογισμικού/ Λογισμικού Ανοιχτού Κώδικα Ελλάδος (ΕΛ/ΛΑΚ)[8].

Οι ψηφιακές βιβλιοθήκες και τα αποθετήρια έχουν φέρει επανάσταση στον τρόπο που αντιμετωπίζουμε και μοιραζόμαστε τις πληροφορίες καθώς επιτρέπουν (Pomerantz & Marchionini, 2007; Candela et al., 2011; Weiss, 2016):

- Ελεύθερη πρόσβαση στις συλλογές επιστημονικού, οικονομικού, νομικού, ιστορικού, πολιτιστικού και ψυχαγωγικού περιεχομένου υιοθετώντας την έννοια της ανοιχτής πρόσβασης στη γνώση για όλους[9].
- Διευκόλυνση του διαμοιρασμού της ανθρώπινης παραγωγής δωρεάν, όπου είναι δυνατό και νόμιμο. Χαρακτηριστικό παράδειγμα είναι τα επιστημονικά άρθρα. Οι εκδοτικοί οίκοι μονοπωλούν την αγορά θέτοντας υπέρογκα ποσά για την απόκτηση ενός άρθρου. Από την άλλη, τα πανεπιστημιακά ιδρύματα απαιτούν την κατάθεση άρθρων (pre-print ή post-print) από το εκπαιδευτικό προσωπικό, ενώ δεν επιβαρύνουν οικονομικά το χρήστη.

[1] Ιστοσελίδα του HathiTrust.

[2] Ιστοσελίδα του Internet Archive.

[3] Ιστοσελίδα της Europeana.

[4] Ιστοσελίδα του PANDORA Archive.

[5] Ιστοσελίδα της World Digital Library.

[6] Ψηφιακή Βιβλιοθήκη Διώνη.

[7] Morgan Museum and Library.

[8] Αποθετήριο ΕΛ/ΛΑΚ.

[9] Τα αποθετήρια επιτρέπουν την ελεύθερη πρόσβαση, ωστόσο σε πολλές περιπτώσεις απαιτείται εγγραφή του χρήστη στο σύστημα.

- Πρόσβαση ανεξάρτητα από χωρικούς και χρονικούς περιορισμούς.
- Πρόσβαση και χρήση των ίδιων ψηφιακών πόρων ταυτόχρονα από πολλούς χρήστες.
- Άμεση και εύκολη ανάκτηση τεκμηρίων.
- Χρήση των αδειών Creative Commons για τον έλεγχο των δικαιωμάτων στα ψηφιακά έργα.
- Δημιουργία και χρήση νέων υποστηρικτικών τεχνολογιών όπως το DSpace και το OAI-PMH protocol.

Αν και η σπουδαιότητά τους είναι φανερή, ανακύπτουν ακόμα πολλά προβλήματα. Κάποια από αυτά αφορούν το χώρο αποθήκευσης των αρχείων, την εκπαίδευση του ανθρωπίνου δυναμικού για τη διαχείρισή τους ή ακόμα και τη χρηματοδότηση για την ομαλή λειτουργία τους.

2.2. Κατηγοριοποίηση/ Ταξινόμηση: Από τον άνθρωπο στη μηχανή

Ο ανθρώπινος εγκέφαλος κάνει συνεχώς και ακούσια κατηγοριοποίηση όλων των πραγμάτων γύρω του. Πρόκειται για μια λειτουργία η οποία αποτελεί τη «βάση για τη δόμηση της γνώσης μας για τον κόσμο» (Cohen & Lefebvre, 2005, σ. 2). Είναι μια «χωρική, χρονική και χωροχρονική κατάτμηση του κόσμου» (Bowker & Star, 2000, σ. 10), που μας επιτρέπει να αντιλαμβανόμαστε, να ξεχωρίζουμε και να αλλάζουμε το περιβάλλον (Puglisi et al., 2008).

Αυτή η διανοητική λειτουργία αποτελεί βασική εργασία για τις φυσικές και τις ψηφιακές βιβλιοθήκες (Hjørland, 2012) και επιτρέπει στους επιστήμονες της πληροφόρησης να ταξινομούν τη γνώση. Ωστόσο, πέρα από τη νοητική ικανότητα, οι ταξινόμοι διαθέτουν εκπαίδευση και γνώση πάνω στη χρήση των κατάλληλων συστημάτων. Συγκεκριμένα, διαχειρίζονται συστήματα θεματικής πρόσβασης (ελεγχόμενα λεξιλόγια όπως είναι οι θεματικές επικεφαλίδες της Εθνικής Βιβλιοθήκης της Ελλάδος, οι θησαυροί EuroVoc και AgroVoc, κ.α.), για τον καθορισμό του θέματος και τη δημιουργία σημείων πρόσβασης, όπως επίσης και ταξινομικά συστήματα.

Η δημιουργία τέτοιων συστημάτων – προτύπων ταξινόμησης γεννήθηκε από την ανάγκη για καθολική κατηγοριοποίηση και ευκολότερη ανάκληση των πληροφοριών. Τα χαρακτηριστικά ενός τέλειου συστήματος ταξινόμησης μπορούν να συνοψιστούν ως (Bowker & Star, 2000; Ruef & Patterson, 2009):

- Συνέπεια και ακολουθία συγκεκριμένων αρχών, όπως τα ιεραρχικά δένδρα.

- Δομημένες κατηγορίες που δεν επιτρέπουν την ταξινόμηση ενός αντικειμένου σε πάνω από μία κατηγορίες.
- Αυτοολοκλήρωση.

Ωστόσο, ένα τέτοιο σύστημα είναι δύσκολο να παραχθεί (Ruef & Patterson, 2009). Έως σήμερα χρησιμοποιούνται παλιά ταξινομικά συστήματα τα οποία μεταβάλλονται συνεχώς με την προσθήκη, μετακίνηση ή κατάργηση κατηγοριών. Από αυτά τα πιο διαδεδομένα παγκοσμίως είναι το Dewey Decimal Classification (DDC)[10], το Library of Congress Classification System (LCC) [11], το Universal Decimal Classification (UDC) [12], το Colon Classification (CC) [13]. Έπειτα υπάρχουν εξειδικευμένα συστήματα, όπως αυτό της National Library of Medicine (NLM)[14], και εθνικά, όπως το Chinese Library Classification (CLC)[15].

Η ταξινόμηση των συστημάτων δεν προκύπτει μόνο από τη χρήση τους. Ανάλογα με τη λειτουργικότητά τους χωρίζονται σε απαριθμητικά, ιεραρχικά, φασετικά ή αναλυτικο-συνθετικά ή πολύπλευρα. Σε κάθε περίπτωση, ο ταξινομικός αριθμός που παράγουν είναι η θέση του τεκμηρίου στο χάρτη της γνώσης όπως αυτός αναπαρίσταται από το εκάστοτε σύστημα.

Εντούτοις, όπως προαναφέρθηκε, η ταξινόμηση είναι μια εργασία που γίνεται σε όλο το φάσμα της ανθρώπινης ζωής και όχι μόνο στις βιβλιοθήκες. Η χρήση του Διαδικτύου απαιτεί την εκμετάλλευση μηχανισμών κατηγοριοποίησης των πόρων στο αχανές σύστημά του. Ένας τέτοιος μηχανισμός είναι η ετικέτα (tag) (Triantafyllou et al., 2014) με την οποία οι χρήστες των blogs, του Twitter, του Facebook, του Instagram και άλλων κοινωνικών δικτύων, είναι εξοικειωμένοι

Μάλιστα, είναι ευρέως διαδεδομένη η χρήση ετικετών τύπου folksonomies. Πρόκειται για ετικέτες ελεύθερου λεξιλογίου που προτείνονται από το σύστημα σύμφωνα με προηγούμενες προτάσεις χρηστών (Specia & Motta, 2007; Schifanella et al., 2010). Με αυτό τον τρόπο αποτυπώνεται και κατηγοριοποιείται εύκολα ένα σύνολο πόρων από τους ίδιους τους δημιουργούς. Το παράδειγμα (Εικόνα 2.2-1) λήφθηκε από

[10] Το DDC παρέχεται από την OCLC.

[11] Το LCC παρέχεται από τη Βιβλιοθήκη του Κογκρέσου.

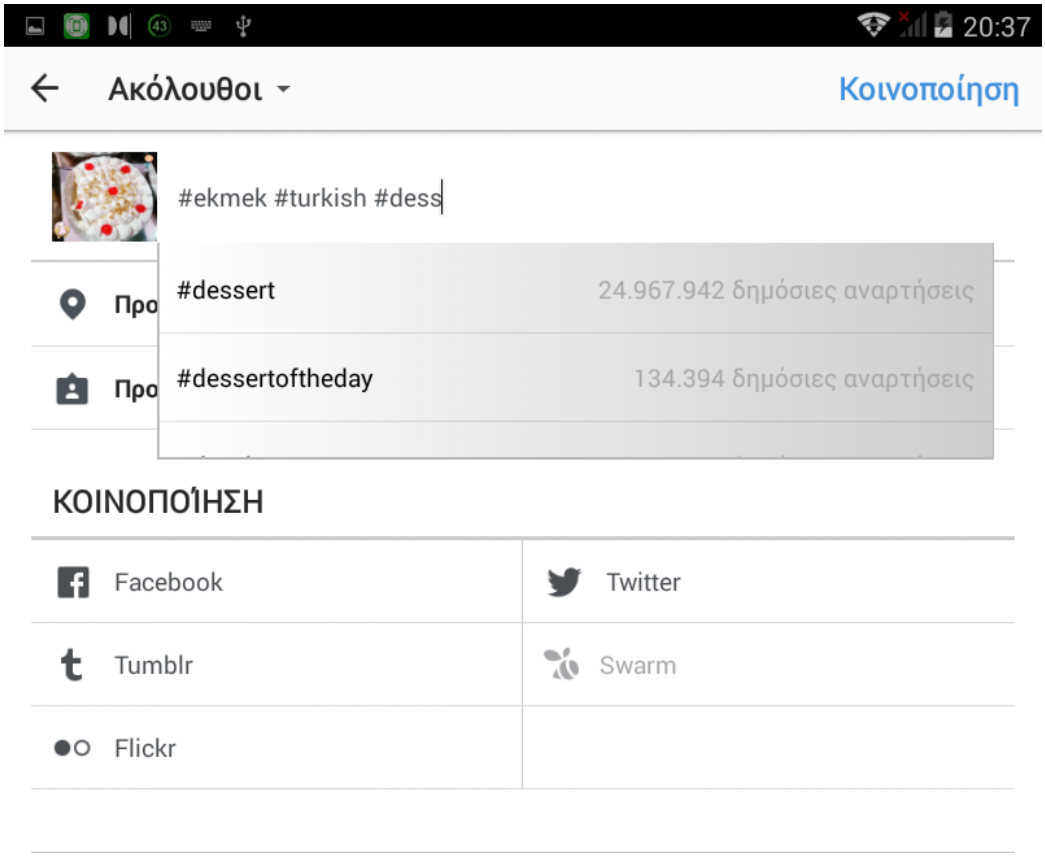
[12] Το UDC παρέχεται από το UDC Consortium.

[13] Το πρώτο αναλυτικο-συνθετικό σύστημα ταξινόμησης από τον S. R. Ranganathan.

[14] Χρησιμοποιείται για την ταξινόμηση στην Ιατρική.

[15] Χρησιμοποιείται σε όλες τις βιβλιοθήκες τις Κίνας, (επιπλέον σύνδεσμος στην Wikipedia).

το Instagram. Όπως φαίνεται, με τη χρήση της δίεσης (#), ή hashtag (Efron, 2010), ο χρήστης επιλέγει όρους – θέματα της φωτογραφίας που θέλει να ανεβάσει. Παράλληλα, το σύστημα προτείνει όρους που έχουν χρησιμοποιηθεί από άλλους χρήστες και παρουσιάζει τον αριθμό των δημοσιεύσεων που περιέχουν το ίδιο hashtag.



Εικόνα 2.2-1. Παράδειγμα folksonomy από το Instagram.

Τέλος, τα τελευταία χρόνια γίνεται εκτενής έρευνα για την ανάπτυξη οντολογιών. Οι οντολογίες στηρίζονται στο Σημασιολογικό Ιστό (Scemantic Web) και αποτυπώνουν πολύπλοκες σχέσεις εννοιών (Staab & Studer, 2013) επιτρέποντας τη διαχείριση της γνώσης (Fensel, 2001; Staab & Studer, 2013).

2.3. Μηχανική μάθηση

2.3.1. Ορισμός και ιστορία

Η Μηχανική Μάθηση (ΜΜ) αποτελεί πεδίο της Τεχνητής Νοημοσύνης (ΤΝ) και υποστηρίζει τους υπολογιστές στη μάθηση χωρίς την ανθρώπινη παρέμβαση (Samuel, 1959). Κατά έναν πιο επίσημο ορισμό «ένα πρόγραμμα υπολογιστή λέγεται ότι

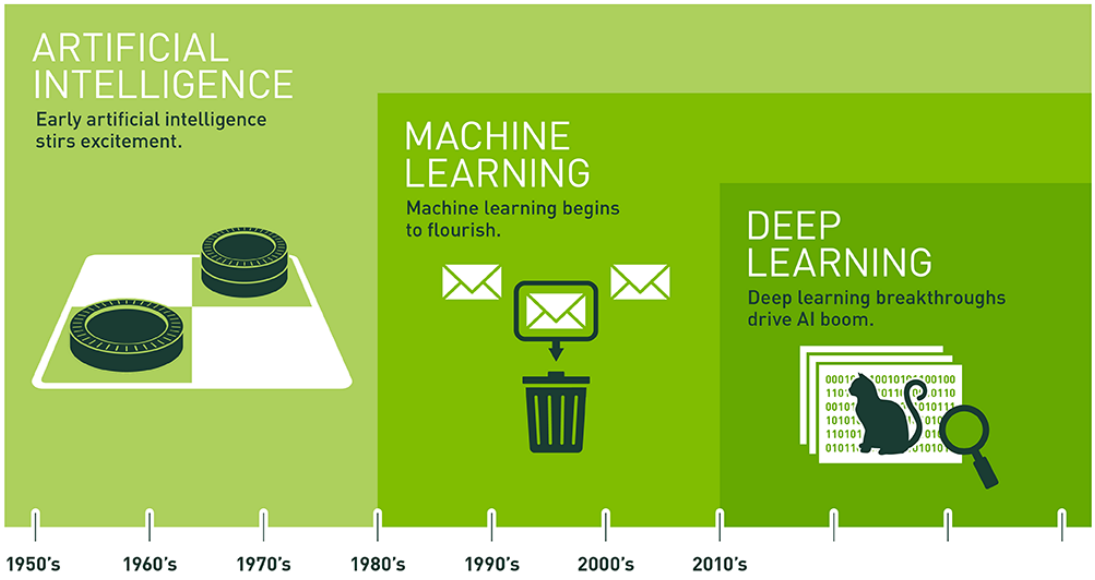
μαθαίνει από μια εμπειρία Ε αναφορικά με μια σειρά από έργα Τ και μια μέτρηση απόδοσης Ρ, εάν η απόδοση του στα έργα Τ, όπως μετριέται από την Ρ, βελτιώνεται με την εμπειρία Ε» (Mitchell, 1997). Η ΜΜ έχει δανειστεί τις μεθόδους και το λεξιλόγιό της από κλάδους όπως τη Στατιστική, τη Λογική και τη Γνωστική Επιστήμη. Η λειτουργία του εγκεφάλου των έμβιων όντων έχει παίξει σημαντικό ρόλο, επιτρέποντας την κατανόηση του τρόπου μάθησης με σκοπό τη μεταφορά του στους υπολογιστές.

Η βασική λογική είναι πως ο υπολογιστής υπάρχει σαν ένα σύστημα που μαθαίνει από τις μεταβολές που συμβαίνουν στη δομή και το πρόγραμμά του ή τις μεταβολές που προκαλούν κάποια εξωτερικά δεδομένα. Το αποτέλεσμα κάθε μεταβολής είναι η αυτοβελτίωση (Nilsson, 1996).

Όπως υποδεικνύεται, ο υπολογιστής διδάσκεται ώστε να φέρει εις πέρας οποιουδήποτε είδους έργο. Όμως, έστω και η υπόνοια αυτής της ιδέας προϋποθέτει την ύπαρξη ευφυΐας. Το 1950, λοιπόν, ο Alan Turing πρότεινε την αντικατάσταση της ευφυούς μηχανής με μια μηχανή που θα μιμείται την ανθρώπινη συμπεριφορά. Έκτοτε η ιδέα οδήγησε στη παραγωγή προγραμμάτων και αλγορίθμων για την επίτευξή της.

Επιγραμματικά, στα χρόνια που ακολούθησαν έγιναν μεγάλα βήματα όπως η δημιουργία του πρώτου προγράμματος μάθησης ντάμας από τον Arthur Samuel το 1952, η δημιουργία του πρώτου νευρωνικού δικτύου (neural network) από τον Frank Rosenblatt το 1957, η ανάπτυξη του αλγορίθμου nearest neighbour (εγγύτερος γείτονας) το 1967 και η σύνθεση πολλών άλλων αλγορίθμων που βαδίζουν στη λογική της ανθρώπινης μάθησης (Marr, 2016).

Αν και η ιστορία της ΤΝ είναι αλληλένδετη με αυτή της ΜΜ, δεν είναι ταυτόσημες (Εικόνα 2.3.1-2). Η πρώτη ασχολείται με την ανάπτυξη συστημάτων μίμησης της ανθρώπινης συμπεριφοράς για την επίλυση πολύπλευρων και πολύπλοκων προβλημάτων και την λήψη αποφάσεων. Από την άλλη, η δεύτερη χρησιμοποιείται σε καθημερινές λειτουργίες, όπως την κατηγοριοποίηση των ειδήσεων, την επεξεργασία δεδομένων, την υποστήριξη και την περάτωση εργασιών (Copeland, 2016).



Εικόνα 2.3.1-2. Σύγκριση της Τεχνητής Νοημοσύνης, της Μηχανικής Μάθησης και του Deep Learning (Copeland, 2016).

2.3.2. Εφαρμογές

Η ΜΜ έχει βρει εφαρμογές σε διάφορους τομείς όπως την Ιατρική Διάγνωση, τη Βιοπληροφορική, την Οικονομία, τη Διαφήμιση, την Επιστήμη των Δεδομένων, την Εξόρυξη των Δεδομένων. Ακόμα έχει εφαρμοστεί στον εντοπισμό ηλεκτρονικής απάτης, τον έλεγχος των ρομπότ, την Οπτική Αναγνώριση Χαρακτήρων (OCR), την αναγνώριση φωνής και εικόνων κ.α. (Wikipedia, 2016a).

Επιπλέον, σημαντική είναι η συνεισφορά της στη κατηγοριοποίηση κειμένου. Η κατηγοριοποίηση κειμένου είναι μια διαδικασία κατά την οποία γίνεται ταξινόμηση εγγράφων σε προκαθορισμένες κατηγορίες με σκοπό την εκπαίδευση ενός αλγορίθμου (Sebastiani, 2002). Μέχρι σήμερα η εφαρμογή της έχει γίνει με επιτυχία, κυρίως με αλγορίθμους τύπου Bayes (Witten et al., 2011), σε εργασίες όπως η κατηγοριοποίηση e-mail (spam ή όχι) και η κατηγοριοποίηση των κορυφαίων θεμάτων στο Twitter (Irani et al., 2010; Awad & ELseuofi, 2011).

2.3.3. Λειτουργίες μάθησης

Το ερώτημα είναι πώς επέρχεται η μάθηση. Υπάρχουν δύο κλασσικές λειτουργίες μάθησης οι οποίες εφαρμόζουν την κατηγοριοποίηση. Η πρώτη ονομάζεται επιτηρούμενη (supervised) μάθηση, ενώ η δεύτερη μη επιτηρούμενη (unsupervised) μάθηση. Έχουν αναπτυχθεί στη λογική της επαγωγικής ή εμπειρικής μάθησης

(inductive/ empirical) (Sammut & Webb, 2011), κατά την οποία ο μαθητευόμενος εξάγει γενικούς κανόνες μέσα από την εμπειρία του με παραδείγματα (Shavlik & Dietterich, 1990).

Κατά την επιτηρούμενη μάθηση, ο αλγόριθμος λαμβάνει δεδομένα μαζί με την υπόδειξη σε ποια κατηγορία ανήκουν. Στόχος είναι να «μαντεύει» σωστά την κατηγορία. Σε αυτήν τη λειτουργία υπάγονται οι διαδικασίες της κατηγοριοποίησης (classification) και της παλινδρόμησης (regression).

Από την άλλη, κατά την μη επιτηρούμενη μάθηση, ο αλγόριθμος καλείται να βρει μόνος του δομή στα δεδομένα εισόδου. Τέτοιες διαδικασίες είναι η συσταδοποίηση (clustering), οι κανόνες συσχέτισης (association rules) και οι χάρτες αυτοοργάνωσης (self-organizing maps).

Τέλος, υπάρχει και μία τρίτη λειτουργία μάθησης που ονομάζεται ενισχυτική μάθηση (reinforcement learning). Ο αλγόριθμος αλληλοεπιδρά με το εξωτερικό περιβάλλον διαμορφώνοντας την κατάλληλη συμπεριφορά για την επίτευξη της εργασίας του. Διαφέρει από την επιτηρούμενη και τη μη επιτηρούμενη μάθηση, διότι, αφενός δεν υπάρχουν εκ των προτέρων ετικέτες, αφετέρου υπάρχει εκ των υστέρων επιβράβευση ανάλογα με την έκβαση της εργασίας (Sammut & Webb, 2011). Στην παρούσα πτυχιακή γίνεται μόνο αναφορά σε αυτή και δεν εξετάζεται περαιτέρω.

2.3.4. Κατηγοριοποίηση (Classification)

Η διαδικασία αυτή έχει σκοπό την ταξινόμηση παραδειγμάτων (ή υποδειγμάτων, ή στιγμών, ή αντικειμένων, ή περιπτώσεων, ή δεδομένων εκπαίδευσης – training data) σε προκαθορισμένες κατηγορίες βάσει των χαρακτηριστικών (ή διανυσμάτων, ή γνωρισμάτων) τους (Sammut & Webb, 2011). Τα παραδείγματα διαθέτουν ετικέτες (labeled data) με την κατηγορία (ή τις κατηγορίες) που ανήκουν, για τη δημιουργία ενός γενικού μοντέλου. Το μοντέλο αξιολογείται σύμφωνα με την απόδοσή του στην ταξινόμηση νέων δεδομένων. Το ποσοστό επιτυχών ταξινομήσεων καθορίζει την ακρίβειά του (Sammut & Webb, 2011).

Μέχρι σήμερα έχουν αναπτυχθεί πολλοί κατηγοριοποιητές οι οποίοι χρησιμοποιούνται σε διάφορες περιπτώσεις. Βασίζονται σε προσεγγίσεις μάθησης και λογικής. Η ποικιλία τους έγκειται στο βασικό κανόνα όπου δεν υπάρχει ένας μόνο κατηγοριοποιητής που να μπορεί να χειριστεί όλες τις περιπτώσεις ταξινόμησης (Charalabopoulos & Anagnostopoulos 2011; Arora & Suman, 2012).

2.3.4.1. Προσεγγίσεις και αλγόριθμοι κατηγοριοποίησης

Πιθανοτικοί κατηγοριοποιητές (Bayesian classifiers)

Ο πιθανοτικοί κατηγοριοποιητές βασίζονται στο ομώνυμο θεώρημα του Τόμας Μπέυζ. Ένας εξ αυτών είναι ο Naive Bayes κατά τον οποίο υπολογίζεται η πιθανότητα ενός εγγράφου να ανήκει σε μια κατηγορία λαμβάνοντας υπόψη τα διανύσματα με τα οποία αναπαρίσταται. Αν και, πρακτικά, τα διανύσματα δεν είναι ανεξάρτητα μεταξύ τους, ο Naive Bayes θεωρεί τις τιμές τους ανεξάρτητες (John & Langley, 1995).

Σύμφωνα με το θεώρημα του Μπέυζ:

$$\mathrm{p}(\mathrm{d}|\mathrm{c}_j) = \frac{p(\mathrm{d}|c_j)p(c_j)}{p(d)} \quad (1)$$

η πιθανότητα του εγγράφου d να ανήκει στην κλάση c_j υπολογίζεται από την εμφάνιση του d στην κλάση c_j επί της συχνότητας της c_j προς της συχνότητας του d. Η απόφαση του κατηγοριοποιητή λαμβάνεται από τον αλγόριθμο:

$$\hat{c} = \arg\max_c P(\mathrm{c}) \prod_{i=1}^{n} P(d_i|c) \quad (2)$$

Αν και θεωρείται ο απλούστερος κατηγοριοποιητής, παράλληλα είναι ο πιο εύχρηστος και ο ταχύτερος (Witten et al, 2011), ενώ έχει εφαρμοσθεί με μεγάλη επιτυχία (Rish, 2001; Kim et al., 2006). Άλλοι πιθανοτικοί κατηγοριοποιητές είναι οι Naive Bayes Multinomial (NBM), Bernoulli Naive Bayes και Gaussian Naive Bayes (McCallum & Nigam, 1998).

Μηχανές διανυσμάτων υποστήριξης (Support vector machines - SVM)

Οι μηχανές διανυσμάτων υποστήριξης ανήκουν στους γραμμικούς κατηγοριοποιητές. Κατά την ταξινόμηση διαχωρίζουν τα παραδείγματα με υπερεπίπεδα αναζητώντας το μέγιστο δυνατό περιθώριο (Sammut & Webb, 2011). Η αποτελεσματικότητα της κατηγοριοποίησης καθορίζεται από το εύρος των περιθωρίων ανάμεσα στα κοντινότερα παραδείγματα και το υπερεπίπεδο. Όσο μεγαλύτερα τα περιθώρια, τόσο ακριβέστερη η ταξινόμηση νέων δεδομένων. Ένα παράδειγμα θα μπορούσε να είναι το παρακάτω. Σύμφωνα με την Εικόνα 2.3.4.1-3, ανάμεσα στα

τρίγωνα και στα τετράγωνα υπάρχουν δύο πιθανά υπερεπίπεδα h_1 και h_2. Αν και τα δύο θα μπορούσαν να εξυπηρετήσουν το σκοπό μας, είναι προφανές ότι το h_1 διαθέτει μεγαλύτερα περιθώρια, άρα είναι το καταλληλότερο.

Το υπερεπίπεδο καθορίζεται από τον τύπο

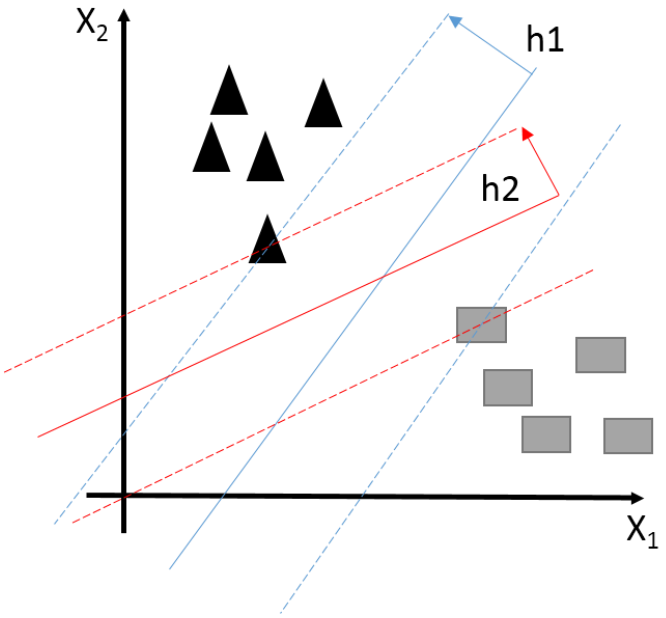
$$g(\vec{x}) = \vec{w}^T \vec{x} + w_0 \tag{3}$$

όπου για κατηγοριοποίηση σε δύο κλάσεις, αν

$$g(\vec{x}) \geq 1 \text{ τότε } \forall \vec{x} \in \text{κλάση τρίγωνα} \tag{4}$$

ή

$$g(\vec{x}) \leq -1 \text{ τότε } \forall \vec{x} \in \text{κλάση τετράγωνα.} \tag{5}$$



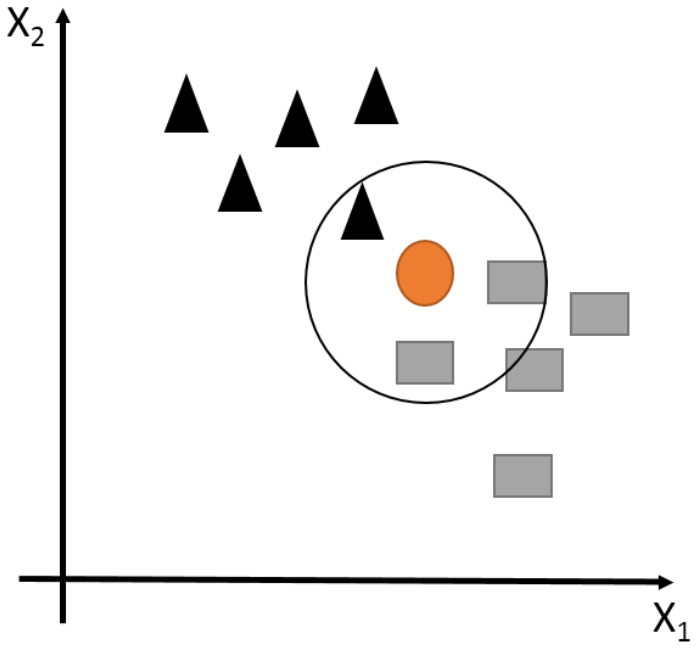
Εικόνα 2.3.4.1-3. Απεικόνιση της κατηγοριοποίησης SVM.

Αν και οι SVM παρουσιάζονται ως γραμμικοί κατηγοριοποιητές, τις τελευταίες δεκαετίες έχει προταθεί η χρήση τους για τη δημιουργία μη γραμμικών κατηγοριοποιητών με τη μέθοδο Kernel (Sammut & Webb, 2011).

Αλγόριθμοι βασισμένοι στα υποδείγματα (Instance-based learning algorithms)

Αυτή η μέθοδος μάθησης περιλαμβάνει αλγορίθμους που βασίζονται σε παλαιότερες κατηγοριοποιήσεις που κρατούν στη μνήμη, ελέγχοντας την ομοιότητά τους (Cleary & Trigg, 1995; Sammut & Webb, 2011). Σε αυτήν την περίπτωση, η κατηγοριοποίηση δεν γίνεται παρά μόνο όταν ζητηθεί στο σύστημα (lazy learning).

Παράδειγμα αποτελεί ο k-Nearest Neighbor (k-NN, όπου το k είναι ο αριθμός των εγγύτερων υποδειγμάτων που θα ληφθούν υπόψη) (Cleary & Trigg, 1995). Για την ταξινόμηση μιας νέας καταχώρησης, υπολογίζεται η απόσταση της από τα ήδη ταξινομημένα παραδείγματα (Peterson, 2009). Η απόφαση βασίζεται στην πλειοψηφία των γειτόνων.



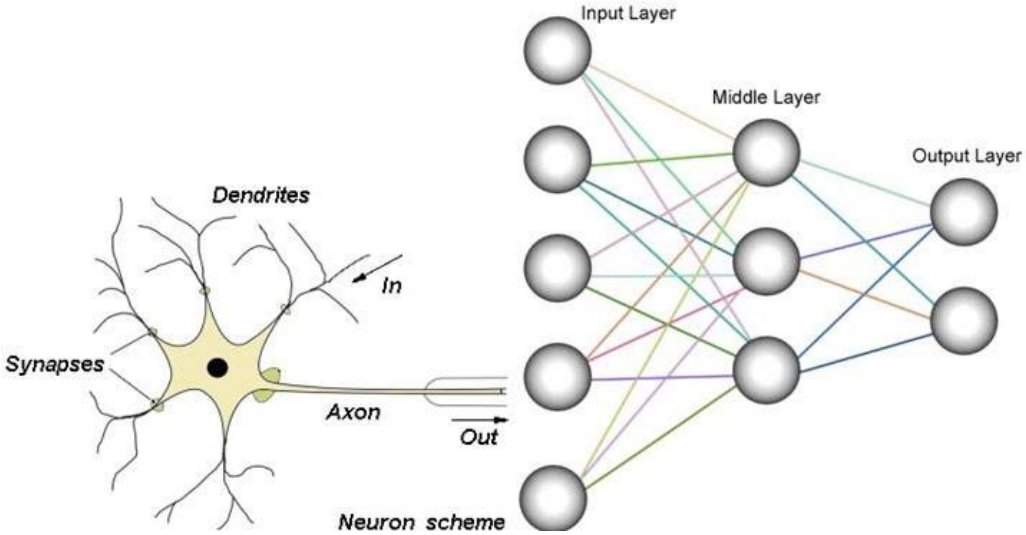
Εικόνα 2.3.4.1-4. Απεικόνιση της κατηγοριοποίησης k-NN.

Αν για k=1, τότε η απόσταση μετριέται ως:

- Ευκλείδεια: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ (6)
- Manhattan: $\sum_{i=1}^{k}|x_i - y_i|$ (7)
- Minkowski: $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$ (8)

Τεχνητά νευρωνικά δίκτυα (Artificial neuron networks)

Τα τεχνητά νευρωνικά δίκτυα βασίζονται στους νευρώνες του νευρικού συστήματος των έμβιων όντων. Ένας νευρώνας είναι κύτταρο με συνδέσεις που του επιτρέπουν να δέχεται και να στέλνει ηλεκτρικά σήματα (Wikipedia, 2016b). Αποτελείται από τους δενδρίτες (dendrites) που συλλέγουν τα ηλεκτρικά σήματα, τις συνάψεις (synapseis) που τα μεταφέρουν στο κύτταρο και αποτελούν τις εισόδους του, το κυρίως μέρος του κυττάρου που αναλαμβάνει όλες τις υπολογιστικές διαδικασίες, και τον άξονα (axon) που στέλνει τα ηλεκτρικά σήματα προς το επόμενο κύτταρο και αποτελεί την έξοδο του (Εικόνα 2.3.4.1-5) (Wikipedia, 2016b). Σε αντίθεση με τους δενδρίτες και τις συνάψεις, ο άξονας κάθε κυττάρου είναι μοναδικός.



Εικόνα 2.3.4.1-5. Σύγκριση φυσικού και τεχνητού νευρώνα[16].

Αντίστοιχα, σε ένα τεχνητό νευρωνικό δίκτυο τα δεδομένα εισέρχονται στους νευρώνες ή κόμβους (nodes) εισόδου, διέρχονται μέσω των συνάψεων και καταλήγουν στους νευρώνες εξόδου. Μεταξύ της εισόδου (x_{ki}) ενός νευρώνα k και της εξόδου y_k μπορεί να υπάρχουν παραπάνω από ένα επίπεδα τα οποία αποτελούνται από τους κρυμμένους νευρώνες όπου είναι υπεύθυνα για την επεξεργασία των δεδομένων. (Abraham, 2005).

Κατά την επεξεργασία τους, τα δεδομένα πολλαπλασιάζονται με το συναπτικό βάρος που τους ισοδυναμεί (w_{ki}), αθροίζονται μεταξύ τους και περνούν από μια συνάρτηση ενεργοποίησης η οποία καθορίζει την τελική τιμή τους.

[16] Πηγή εικόνας: https://sites.google.com/site/mrstevensonstechclassroom/hl-topics-only/4a-robotics-ai/neural-networks-computational-intelligence

$$y_k = f \sum_{i=0}^{N} x_{ki} w_{ki} \quad (9)$$

Η συνάρτηση ενεργοποίησης μπορεί να είναι:

- Βηματική: $f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ (10)
- Γραμμική: $f(x) = x$ (11)
- Σιγμοειδής: $f(x) = \frac{1}{1+e^x}$. (12)

Τέλος, τα δεδομένα στέλνονται στους επόμενους κόμβους, επαναλαμβάνοντας την διαδικασία επεξεργασίας μέχρι να καταλήξουν στους νευρώνες εξόδου (Abraham, 2005).

Ένας αλγόριθμος που ανήκει στην κατηγορία των νευρωνικών δικτύων είναι ο Multilayer Perceptron. Ανάμεσα στα χαρακτηριστικά του είναι ότι διαθέτει πολλά κρυμμένα επίπεδα νευρώνων, ακολουθεί την εμπροσθόδρομη τροφοδοσία και τον κανόνα οπισθοδιάδοσης του λάθους (backpropagation) (Sammut & Webb, 2011).

Δένδρα απόφασης (Decision trees)

Αυτοί οι αλγόριθμοι βασίζονται στις ομώνυμες δενδροειδής δομές. Ξεκινούν συνήθως από πάνω προς τα κάτω, με τον πρώτο κόμβο να αποτελεί τη ρίζα (root). Κάθε κόμβος συνδέεται με τον προηγούμενο και τον επόμενο με κλαδιά (branches) μέχρι να καταλήξουν στο τελικό κόμβο που ονομάζεται φύλλο (leaf) (Witten et al., 2011).

Ο βασικός αλγόριθμος που χρησιμοποιείται ονομάζεται Top-Down Induction of Decision Trees (TDIDT) (εναλλακτικά recursive partitioning ή «διαίρει και βασίλευε») (εκτενής αναφορά στο Quinlan, 1986). Κατά αυτόν, ως ρίζα επιλέγεται το καταλληλότερο χαρακτηριστικό, προσθέτονται τα κλαδιά και οι κόμβοι, ενώ παράλληλα διαιρούνται τα παραδείγματα (Sammut & Webb, 2011).

Σημαντικό ρόλο έχει η επιλογή των χαρακτηριστικών για τα οποία εφαρμόζονται δύο δείκτες νοθείας (impurity measures), ο information-theoretic entropy

$$Entropy(S) = -\sum_{i=1}^{c} \frac{|S_I|}{|S|} * log_2 \left(\frac{|S_i|}{S} \right) \quad (13)$$

και ο Gini Index

$$Gini(S) = 1 - \sum_{i=1}^{c} \left(\frac{|S_I|}{|S|} \right)^2 \quad (14)$$

όπου το S είναι το σύνολο των δεδομένων εκπαίδευσης και S_i είναι το υποσύνολο που ανήκει σε μια κατηγορία (Breinman et al., 1984; Quinlan, 1986; Sammut & Webb, 2011).

Κάποιοι κατηγοριοποιητές αυτής της προσέγγισης είναι ο Random Forest, ο Random Tree και ο J48.

Rule based learners

Σε αυτή την οικογένεια αλγορίθμων, η βασική λογική είναι κατηγοριοποίηση μέσω εκμάθησης κανόνων. Κυριαρχεί η προσέγγιση του «διαίρει και βασίλευε» όπου από το εκάστοτε σύνολο εκπαίδευσης διαιρείται σε κομμάτια πάνω στα οποία ο αλγόριθμος καλείται να βρει τους κατάλληλους κανόνες. Αυτό συμβαίνει μέχρι να μην υπάρχουν άλλα παραδείγματα και κάθε παράδειγμα να αντιστοιχεί σε έναν κανόνα (Fürnkranz, 1999). Παραδείγματα κατηγοριοποιητών είναι ο PART και ο JRip.

Metalearners

Σε αυτή την προσέγγιση, η τελική απόφαση ταξινόμησης λαμβάνεται από συνδυασμό αποφάσεων πολλών ασθενών μοντέλων (weak learners). Οι κύριες μεθοδολογίες που χρησιμοποιούνται βασίζονται στα μοντέλα μάθησης τύπου ensemble, Bagging, Boosting και Stacking (Sammut & Webb, 2011). Τέτοιοι ταξινομητές είναι ο AdaBoost και ο LogitBoost.

2.3.5. Συσταδοποίηση (Clustering)

Σε αντίθεση με την κατηγοριοποίηση, η συσταδοποίηση καλείται να κατηγοριοποιήσει παραδείγματα χωρίς να είναι γνωστή η κατηγορία στην οποία ανήκουν (unlabeled) (Alpaydin, 2010). Οι συστάδες (συγκεντρωμένο σύνολο = cluster) που δημιουργούνται αποτελούνται από παραδείγματα με κοινά χαρακτηριστικά. Μια συσταδοποίηση θεωρείται επιτυχημένη όταν μια συστάδα παρουσιάζει μεγάλη ομοιογένεια ανάμεσα στα παραδείγματά της και μεγάλη απόκλιση ή διαφορά από άλλες συστάδες (Alpaydin, 2010).

Η συσταδοποίηση γίνεται βάσει μοντέλων με συγκεκριμένα χαρακτηριστικά τα οποία προσδιορίζουν το είδος της. Τα είδη της μπορούν να την καθορίσουν ως (Sammut & Webb, 2011; Witten et al., 2011):

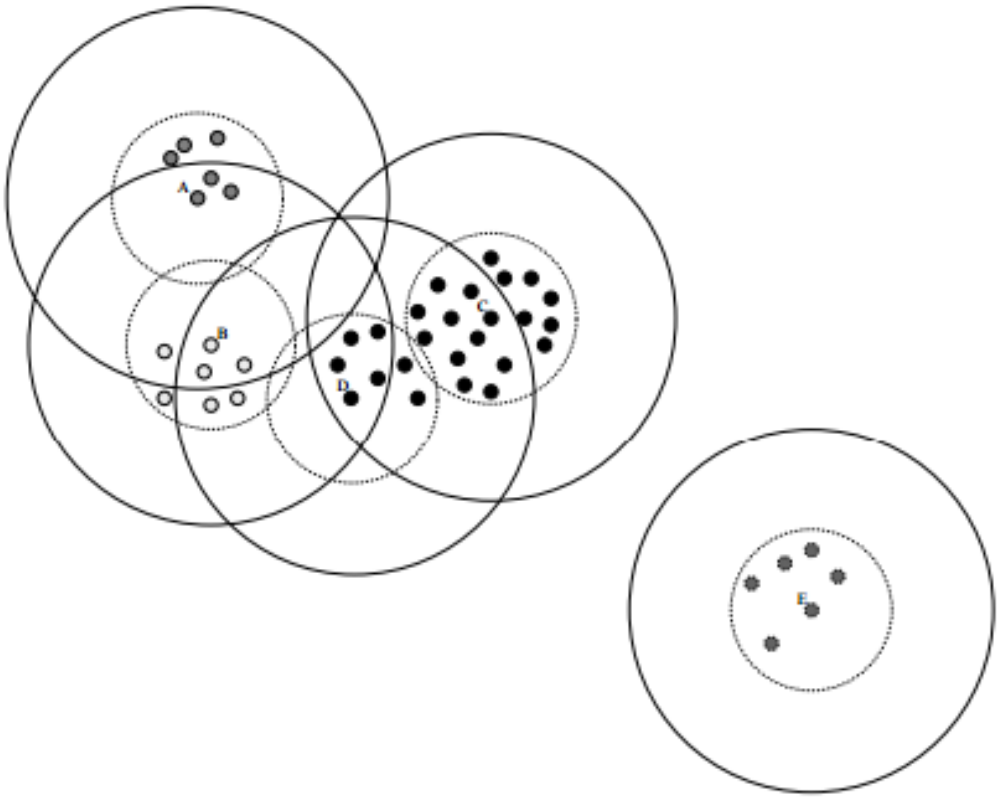
- Ιεραρχική (hierchical): Επιτρέπει τη δημιουργία φωλιασμένων συστάδων που εμφανίζονται με δενδροειδή μορφή.
- Διαχωριστική (partitional): Διαίρεση των συστάδων ώστε κάθε περίπτωση να υπάρχει σε μία μόνο συστάδα.
- Αποκλειστική (exclusive): Κάθε περίπτωση ανήκει σε μία μόνο συστάδα.
- Επικαλυπτόμενη (overlapping): Οι περιπτώσεις μπορούν να ανήκουν ταυτόχρονα σε παραπάνω από μία κατηγορίες.
- Ασαφής (fuzzy): Κάθε περίπτωση ανήκει σε όλες τις συστάδες τόσο όσο δείχνει το βάρος του. Άρα για το βάρος 0-1, όπου 0, δεν ανήκει καθόλου, ενώ 1, ανήκει τελείως, το άθροισμα του βάρους από όλες τις συστάδες πρέπει να ισούται με 1 για κάθε μια περίπτωση.
- Πλήρης (complete): Γίνεται συσταδοποίηση σε όλες τις περιπτώσεις.
- Επιμέρους (partial): Κάποιες περιπτώσεις δεν εκχωρούν σε συστάδες.

2.3.5.1. Τεχνικές συσταδοποίησης

Canopy

Η μέθοδος Canopy υποστηρίζει την ομαδοποίηση σε συστάδες. Μπορεί να χρησιμοποιηθεί είτε μόνη της είτε ως προπαρασκευαστικό στάδιο σε τεχνικές K-Means και ιεραρχικούς clusterers για μεγάλα σύνολα δεδομένων. Βασίζεται στο διαχωρισμό υποσυνόλων (canopies), τα οποία δημιουργούνται με την τυχαία επιλογή ενός

αντικειμένου και τη χρήση μιας μετρικής εκτίμησης απόστασης (approximate distance measure) (McCallum et al., 2000). Επιπλέον, θέτονται δύο όρια γύρω από το αρχικό αντικείμενο, T1>T2. Για κάθε νέο παράδειγμα, αν η απόστασή του είναι <T1, μπαίνει στην ίδια συστάδα (Εικόνα 2.3.5.1-6), αλλιώς, αν η απόστασή του είναι μεταξύ T1 και T2, επαναλαμβάνεται η διαδικασία (McCallum et al., 2000).

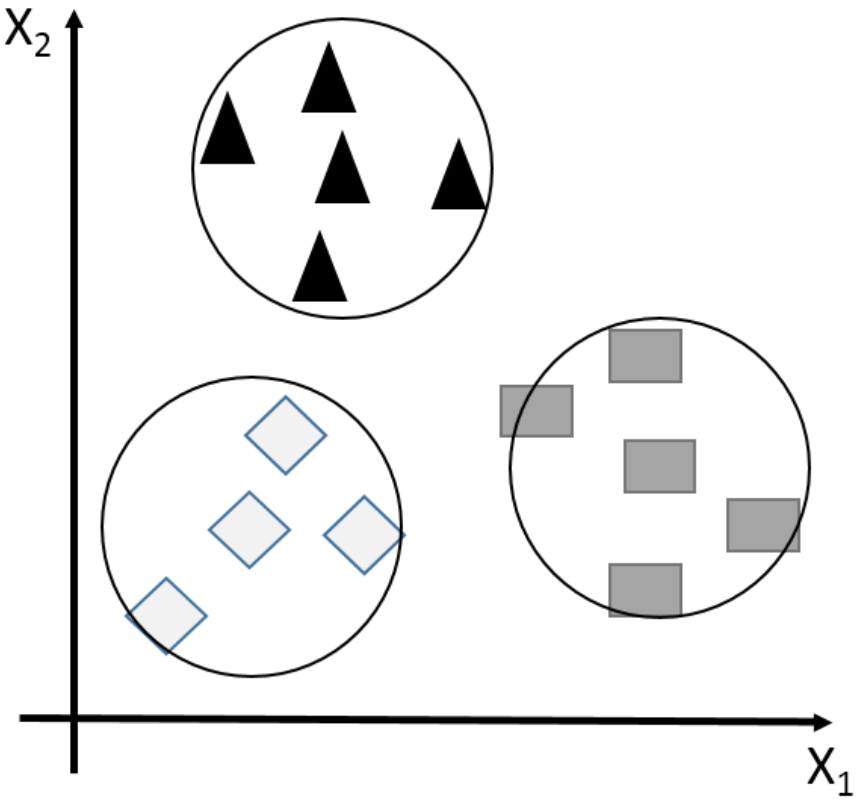


Εικόνα 2.3.5.1-6. Παράδειγμα Canopy με επικαλυπτόμενες συστάδες, οι οποίες προκύπτουν από το μεγάλο μέγεθος των ορίων (McCallum et al., 2000).

Κεντροειδής συσταδοποίηση (Centroid-based)/ αλγόριθμος k-means

Ο αλγόριθμος k-means βασίζεται στην εύρεση τυχαίων σημείων, k στον αριθμό, που αποτελούν τα κέντρα των συστάδων. Για κάθε παράδειγμα, ο αλγόριθμος υπολογίζει την απόστασή του από τα κέντρα των συστάδων (ευκλείδεια απόσταση). Η τοποθέτησή του γίνεται στη συστάδα με το κοντινότερο προς αυτό κέντρο (Εικόνα 2.3.5.1-7). Έπειτα, λαμβάνεται ο μέσος όρος των παραδειγμάτων κάθε συστάδας και η διαδικασία επαναλαμβάνεται με την αλλαγή των κέντρων (Witten et al., 2011).

Ο k-means είναι από τους πιο δημοφιλής αλγορίθμους για τη συσταδοποίηση, ενώ έχουν δημιουργηθεί πολλές παραλλαγές του, όπως ο Cascade Simple K-Means και ο X-Means.



Εικόνα 2.3.5.1-7. Απεικόνιση συσταδοποίησης k-means.

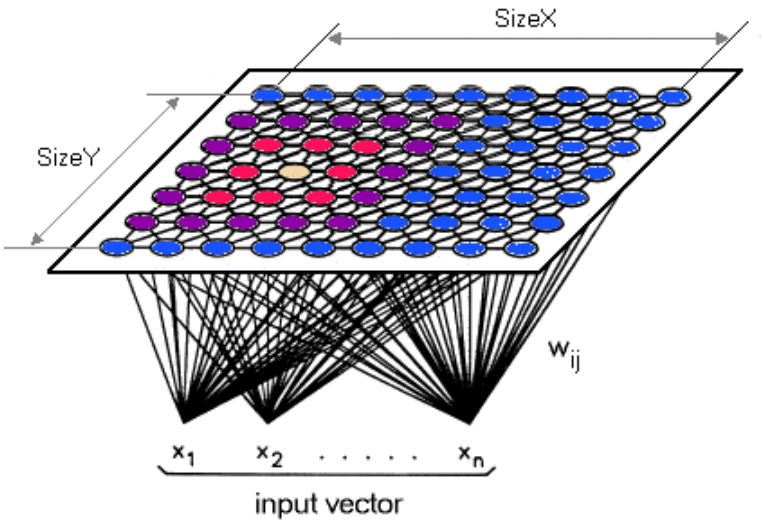
Συσταδοποίηση που βασίζεται στην κατανομή (Distribution-based)/ Ο αλγόριθμος Expectation Maximization

Ο αλγόριθμος Expectation Maximization (EM) χρησιμοποιείται συχνά όταν τα δεδομένα είναι ελλιπή. Είναι μέθοδος μεγιστοποίησης της πιθανοφάνειας όπου κάθε αντικείμενο λαμβάνει μια πιθανότητα να ανήκει σε κάθε μια από τις συστάδες.

Η διαδικασία που ακολουθεί, χωρίζεται σε δύο φάσεις. Στην πρώτη, εκτιμάται η μέση τιμή της συνάρτησης πιθανότητας για το σύνολο των δεδομένων, βάσει των τιμών των γνωστών και των υπό εκτίμηση παραμέτρων (expectation), ενώ στη δεύτερη υπολογίζονται οι τιμές των παραμέτρων που αυξάνουν τη μέση τιμή (maximization). Η διαδικασία επαναλαμβάνεται μέχρι την εύρεση της μέγιστης πιθανότητας (Peel & McLachlan, 2000; Do & Batzoglou, 2008).

Χάρτες αυτοοργάνωσης (Self-organizing maps)

Οι χάρτες αυτοοργάνωσης (Self-organizing maps - SOM) ή χάρτες Kohonen είναι ένας τύπος τεχνητών νευρωνικών δικτύων. Χρησιμοποιούνται στη μη επιβλεπόμενη μάθηση για τη δημιουργία πλεγμάτων συστάδων όπου απεικονίζονται δεδομένα σε χαμηλές διαστάσεις (συνήθως δισδιάστατα) στο επίπεδο εξόδου (Sammut & Webb, 2011). Ο SOM αποτελείται από νευρώνες οι οποίοι ανταγωνίζονται μεταξύ τους για την ενεργοποίησή τους σε συγκεκριμένα σήματα εισόδου. Ο νευρώνας που ενεργοποιείται είναι ο νικητής, ο οποίος ορίζει τη θέση των κοντινών διεγερμένων νευρώνων. Οι διεγερμένοι νευρώνες, με τη σειρά τους, ανταποκρίνονται αυξάνοντας τις τιμές τους και τροποποιώντας τις τιμές των συναπτικών βαρών.



Εικόνα 2.3.5.1-8. Απεικόνιση της συσταδοποίησης με SOM[17].

[17] Πηγή εικόνας: https://codesachin.wordpress.com/2015/11/28/self-organizing-maps-with-googles-tensorflow/

2.4. Διαχείριση δεδομένων

Αν και ο λόγος είναι θεμέλιο της ανθρώπινης επικοινωνίας, ένας αλγόριθμος δεν αντιλαμβάνεται το νόημα του και οι λέξεις δεν σημαίνουν κάτι. Ένα κείμενο πρέπει να αναλυθεί σε σημείο όπου να βρεθούν τα όρια των προτάσεων, να διαχωρισθούν οι λεκτικές μονάδες και άλλα στοιχεία όπως ημερομηνίες και ακρωνύμια, να γίνει εκκαθάριση (μείωση του θορύβου), μετατροπή και αναπαράσταση σε αναλύσιμη μορφή. Η διαδικασία γίνεται με τυποποιημένο τρόπο με τη βοήθεια ενός text handler (Triantafyllou et al., 2001) και σκοπός της είναι να τροφοδοτήσει με κατανοητά δεδομένα τον αλγόριθμο.

2.4.1. Τετριμμένες/ Λειτουργικές λέξεις

Οι λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης στο λόγο δεν έχουν ιδιαίτερο νόημα. Αν και βοηθούν την ανθρώπινη επικοινωνία, επιβαρύνουν τη διαδικασία εντοπισμού της ορολογίας με αποτέλεσμα να πρέπει να απομακρυνθούν. Στη βιβλιογραφία αναφέρονται ως stop words, ή στα ελληνικά τετριμμένες/ λειτουργικές λέξεις (Wilbur & Sirotkin, 1992). Ανάλογα με την περίπτωση που εξετάζεται, δημιουργούνται διαφορετικές λίστες με τετριμμένες/ λειτουργικές λέξεις, ενώ δεν έχουν εκδοθεί επίσημες λίστες για κάθε γλώσσα. Συνήθως στα ελληνικά είναι σύνδεσμοι, άρθρα, προθέσεις, αντωνυμίες, ή ακόμα και ρήματα όπως το «είναι», «έχει», «πρέπει».

2.4.2. Word stemming (Στατιστική αναγνώριση λήμματος)

Η ελληνική γλώσσα είναι ιδιαίτερα κλητική, γεγονός που σημαίνει ότι η ίδια λέξη μπορεί να εμφανίζεται στο ίδιο κείμενο με διαφορετικές καταλήξεις. Η διαδικασία word stemming (στατιστική αναγνώριση λήμματος) αφορά την μείωση μιας λέξης στο θέμα της και την εύρεση των μορφολογικών παραλλαγών της (Sembok & Bakar, 2011).

Ένας αλγόριθμος συγκεντρώνει όμοιες υποψήφιες λέξεις θέτοντάς τες σε ένα λήμμα υπό έναν αντιπροσωπευτικό όρο και τις κατατάσσει ανάλογα με την ομοιότητά τους σύμφωνα με ένα μηχανισμό εκτίμησης ομοιότητας που βασίζεται στα N-γράμματα (n-grams). Η μέθοδος n-gram επιτρέπει τη σύγκριση των λέξεων ανά γράμμα

(n), δίνοντας τη δυνατότητα στο χρήστη να θέσει ανά πόσα γράμματα θα γίνεται (1-gram, 2-grams, 3grams, κ.λπ.) (Cavnar & Trenkle, 1994).

2.5. Μέτρηση βαρύτητας όρων

Η μέτρηση βαρύτητας ενός όρου δείχνει τη σημασία του μέσα σε ένα έγγραφο ή ένα σώμα κειμένων. Το αποτέλεσμα της διαδικασίας παράγει μια λίστα η οποία κατατάσσει υψηλότερα τις λέξεις με αυξημένη σημασία και εξαρτάται από τη μέθοδο που έχει επιλεγεί (Jones, 1972; Croft et al., 2010).

Η TF.IDF (term frequency * inverse document frequency) είναι κλασσική μέθοδος ενώ παρουσιάζεται με πολλές παραλλαγές. Σύμφωνα με μια από αυτές το TF αντιπροσωπεύει τον αριθμό των φορών που εμφανίζεται ένας όρος f στο κείμενο, ενώ το IDF είναι ο λογάριθμος του συνολικού αριθμού των εγγράφων, N, προς τον αριθμό των εγγράφων που περιέχουν τον όρο f, N_f. Αυτό συμβαίνει διότι το TF κάνει όλους τους όρους σημαντικούς και αυξάνει το σκορ των τετριμμένων λέξεων, ενώ το IDF τους μετριάζει, προωθώντας το σκορ των λιγότερο τακτικών λέξεων.

$$\mathrm{TF.IDF} = \mathrm{tf} \,.\, \log \frac{N}{N_f} \tag{15}$$

Αν και έχει χρησιμοποιηθεί περισσότερο από άλλες μεθόδους, η TF.IDF έχει γίνει αντικείμενο κριτικής, ενώ κατά καιρούς έχουν προταθεί μετρικές (Yang & Liu, 1999; Soucy & Mineau, 2005; Lan et al., 2006; Lan et al., 2009;) για να την αντικαταστήσουν (Ramos, 2003; Chelba et al., 2008).

2.6. WEKA

Το λογισμικό WEKA αποτελεί μια ολοκληρωμένη λύση για τους χρήστες των αλγορίθμων μηχανικής μάθησης. Αν και προωθείται, κυρίως, σαν λογισμικό εξόρυξης δεδομένων, παρέχει εργαλεία που χρησιμοποιούνται ευρέως (Bouckaert et al., 2010). Οι εφαρμογές του επιτρέπουν την προ-επεξεργασία δεδομένων, την κατηγοριοποίηση, την παλινδρόμηση, τη συσταδοποίηση, τους κανόνες συσχέτισης, την οπτικοποίηση δεδομένων και την ανάπτυξη αλγορίθμων (Bouckaert et al., 2010; Machine Learning Group at the University of Waikato, 2016).

Δημιουργήθηκε από την Ομάδας Μηχανικής Μάθησης του Πανεπιστημίου του Waikato της Νέας Ζηλανδίας. Το λογισμικό ξεκίνησε να αναπτύσσεται το 1993 γραμμένο σε γλώσσα προγραμματισμού C. Συγκριτικά με την τρέχουσα έκδοση, είχε ελάχιστες δυνατότητες. Αργότερα, το 1997, αποφασίσθηκε η αναδημιουργία του σε γλώσσα προγραμματισμού Java, η οποία χρησιμοποιείται έως σήμερα (Bouckaert et al., 2010).

Εικόνα 2.6-9. Το λογότυπο του WEKA.

Το WEKA είναι λογισμικό ανοικτού κώδικα και διανέμεται με GNU General Public License. Λειτουργεί στις πλατφόρμες Windows, Mac OS X και Linux. Η τρέχουσα stable έκδοση είναι η 3.8 ενώ η 3.9 διατίθεται για developers. Όλες οι προηγούμενες εκδόσεις είναι διαθέσιμες στο SourceForge[18].

[18] Όλες οι εκδόσεις του WEKA.

Χαρακτηριστικό της δημοτικότητάς του είναι οι δράσεις που γίνονται για την εκμάθηση του. Συγκεκριμένα παρέχεται ένα Wikispace[19], ένα Massive Open Online Course (MOOC)[20] στο YouTube και το βιβλίο "Data Mining: Practical Machine Learning Tools and Techniques"[21] από τους προγραμματιστές του.

Ωστόσο, όπως έχει ήδη αναφερθεί, το WEKA είναι λογισμικό ανοικτού κώδικα, άρα η ανάπτυξή του δεν υποστηρίζεται μόνο από την Ομάδα Μηχανικής Μάθησης. Ενδεικτικά, μια απλή αναζήτηση στο αποθετήριο GitHub[22] επιστρέφει ένα μεγάλο αριθμό αποτελεσμάτων. Ανάμεσα σε αυτά υπάρχουν ολόκληρα αποθετήρια, πολλές γραμμές κώδικα και συνομιλίες μεταξύ ελεύθερων προγραμματιστών. Επιπλέον, η κοινότητα προγραμματιστών του Stack Overflow[23] αναρτά συζητήσεις με σκοπό την αλληλοβοήθεια, την επίλυση θεμάτων και τη βελτίωση του WEKA.

Από το 2006 το λογισμικό υποστηρίζεται οικονομικά από την εταιρία Pentaho, η οποία εκδίδει το ομώνυμο λογισμικό ανοικτού κώδικα [24] εκμετάλλευσης της επιχειρηματικής ευφυΐας. Συγχρόνως, το WEKA έχει συνδεθεί με το προαναφερθέν λογισμικό για την εκτέλεση εργασιών εξόρυξης δεδομένων (Hall et al., 2009; Bouckaert et al., 2010).

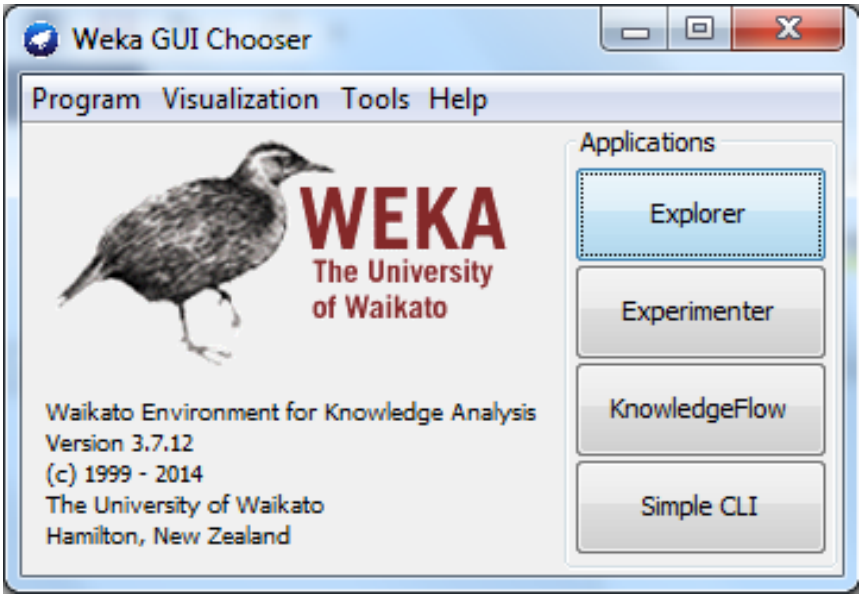
2.6.1. Το περιβάλλον του WEKA

Κατά την εκκίνηση του προγράμματος εμφανίζεται το παράθυρο επιλογής (WEKA GUI Chooser) (Εικόνα 2.6.1-10). Από εκεί ο χρήστης μπορεί να επιλέξει μια από τις τέσσερις εφαρμογές Explorer, Experimenter, Knowledge Flow και Simple CLI:

- Ο Explorer δίνει τη δυνατότητα διαχείρισης ενός συνόλου δεδομένων. Υπάρχει επιλογή προεπεξεργασίας και οπτικοποίησης των δεδομένων, εφαρμογής αλγορίθμων classification, clustering και association rules.
- Ο Experimenter επιτρέπει την προσομοίωση πολλών αλγορίθμων σε πολλά σύνολα δεδομένων. Δεν υπάρχει επιλογή προεπεξεργασίας και οπτικοποίησης των δεδομένων.

[19] http://weka.wikispaces.com/.
[20] WEKA MOOC.
[21] Data Mining: Practical Machine Learning Tools and Techniques.
[22] Αποτελέσματα αναζήτησης GitHub.
[23] Ερωτήσεις στο Stack Overflow.
[24] Pentaho.

- Η εφαρμογή Knowledge Flow παρέχει σχεδόν τις ίδιες δυνατότητες με τον Explorer, με τη διαφορά ότι γίνεται οπτική παρουσίαση των εργασιών με τη μορφή ροής δεδομένων.
- Η Simple CLI επιτρέπει την επικοινωνία με το σύστημα και την εκτέλεση εργασιών μέσω της γραμμής εντολών.



Εικόνα 2.6.1-10. Η διεπαφή του WEKA.

2.6.2. Εφαρμογές του WEKA

Μέχρι σήμερα έχουν δημοσιευθεί πολλά άρθρα στα οποία γίνονται αναφορές για την εφαρμογή του λογισμικού. Φυσικά, τα θέματα διαφέρουν, γεγονός που δείχνει την επιτυχία του σε ένα μεγάλο φάσμα προβλημάτων όπως στην πρόβλεψη μελλοντικών επιλογών (Bresfelean, 2007), στην εξαγωγή αποτελεσμάτων από ερωτηματολόγια (Rajput et al., 2011), το μάρκετινγκ (Turcinek et al., 2012) κ.α.

Η βιοιατρική είναι ένας ακόμα τομέας που έχει εκμεταλλευθεί τις δυνατότητες του WEKA (Frank et al., 2004). Οι Bin Othman & Yau (2007) χρησιμοποίησαν το WEKA και τις μεθόδους Naive Bayes Network, Radial Basis Function, Decision Tree, Single Conjunctive Rule Learner και Nearest Neighbor για την ανίχνευση του καρκίνου του μαστού. Το δείγμα τους αποτελούνταν από 699 περιπτώσεις των 9 διανυσμάτων. Σύμφωνα με τα αποτελέσματα ο Naive Bayes Network απέδωσε την καλύτερη κατηγοριοποίηση από όλους.

Οι Charalabopoulos & Anagnostopoulos (2011) εξέτασαν τη δυνατότητα ταξινόμησης ιστοσελίδων με αλγόριθμους κατηγοριοποίησης και συσταδοποίησης. Το δείγμα χωρίστηκε σε δυο ομάδες των 100 και 1000 περιπτώσεων, με κάθε περίπτωση να περιγράφεται από 10 γνωρίσματα σχετικά με το θέμα και 1 γνώρισμα που αφορά την κλάση (Charalabopoulos & Anagnostopoulos, 2011). Χρησιμοποιήθηκαν οι κατηγοριοποιητές Multilayer Perceptron, M5 Rules, REPTree, k-NN, και οι συσταδοποιητές Farthest Fisrt, SimpleKMeans, SimpleEM. Σύμφωνα με τα αποτελέσματα, η συσταδοποίηση απέδωσε καλύτερα από την κατηγοριοποίηση, ενώ τα δένδρα απόφασης παρήγαγαν καλύτερα και πιο γρήγορα αποτελέσματα από το Multilayer Perceptron (Charalabopoulos & Anagnostopoulos, 2011).

Οι Joshi & Panchal (2014) χρησιμοποίησαν το WEKA με σκοπό την ανάλυση των υποψήφιων για τις θέσεις επίκουρου καθηγητή και αναπληρωτή καθηγητή του Πανεπιστημίου Grant Commission. Η απόφαση αποδοχής σε μια από τις δυο θέσεις, ή απόρριψης, έγινε σύμφωνα με τον Ακαδημαϊκό Δείκτη Απόδοσης. Ο αλγόριθμος που εξετάστηκε ήταν το δένδρο αποφάσεων J48. Τα αποτελέσματα λήφθηκαν υπόψη της Διοίκησης Ανθρωπίνου Δυναμικού.

Τέλος, το 2014 οι Yadav et al. με τη χρήση του WEKA και αλγορίθμων τύπου artificial neuron network (ANN) κατάφεραν να εντοπίσουν τις κατάλληλες μεταβλητές για την πρόβλεψη μοντέλων ηλιακής ακτινοβολίας. Τα αποτελέσματά τους συνεισφέρουν στον τομέα της εκμετάλλευσης ανανεώσιμων πηγών ενέργειας και άλλων συναφών με το θέμα.

3. Μεθοδολογία

3.1. Συλλογή δεδομένων

Το πρώτο στάδιο της έρευνας ήταν η συλλογή των δεδομένων. Στόχος ήταν η συγκέντρωση εγγραφών από τις οποίες θα εξάγονταν οι περιλήψεις. Συνολικά, συγκεντρώθηκαν 718 εγγραφές διατριβών και επιστημονικών άρθρων. Οι εργασίες είχαν ήδη ταξινομηθεί σε μία από τις κατηγορίες Ιατρική, Τουρισμός, Τεχνολογία Τροφίμων, οι οποίες μετά από αναζήτηση, αποδείχτηκαν οι πολυπληθέστερες. Η αναζήτηση έγινε στις ψηφιακές βιβλιοθήκες και τα αποθετήρια 9 ακαδημαϊκών ιδρυμάτων:

- Υπατία- Τεχνολογικό Εκπαιδευτικό Ίδρυμα Αθηνών (512),
- Ψηφιακό αποθετήριο του Γεωπονικού Πανεπιστημίου Αθηνών (73),
- Εύρηκα!- Τεχνολογικό Ίδρυμα Θεσσαλονίκης (47),
- Διώνη- Πανεπιστήμιο Πειραιά (45),
- Ψηφίδα- Πανεπιστήμιο Μακεδονίας (19),
- DSpace@NTUA- Εθνικό Μετσόβιο Πολυτεχνείο (11),
- Νημερτής- Πανεπιστήμιο Πατρών (9),
- E-Locus- Πανεπιστήμιο Κρήτης (1),
- Ανάκτησις- Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Μακεδονίας (1).

Πρέπει να σημειωθεί ότι τέτοιες έρευνες απαιτούν ένα μεγάλο σώμα κειμένων. Αν και οι περισσότερες περιλήψεις λήφθηκαν από την Υπατία, ήταν απαραίτητη η συνδρομή και των υπόλοιπων βιβλιοθηκών. Την περίοδο της συλλογής δεδομένων η Υπατία ήταν ακόμα σε διαδικασία εμπλουτισμού άρα δεν μπορούσε να παρέχει εξολοκλήρου το απαιτούμενο υλικό.

Ωστόσο, κάθε βιβλιοθήκη χρησιμοποιεί διαφορετικά συστήματα ταξινόμησης και θεματικής πρόσβασης. Κάποιες, μάλιστα, επιλέγουν να μην εμφανίσουν την ακριβή θέση του τεκμηρίου στο σύστημα ταξινόμησης, παρά μόνο τα αποτελέσματα της θεματικής ανάλυσης. Για την επίτευξη ομοιογένειας και συμφωνίας εντός του συνόλου των δεδομένων, χρησιμοποιήθηκε το Δεκαδικό Σύστημα Ταξινόμησης Dewey το οποίο καθόρισε την κατηγορία που υπάγονται τα άρθρα και, κατ' επέκταση, αν θα συμπεριλαμβάνονταν ή θα απορρίπτονταν. Τη μόνη εξαίρεση αποτέλεσε ένα σύνολο 22 περιλήψεων από διπλωματικές του τμήματος Επιστήμης και Τεχνολογίας Τροφίμων

οι οποίες κρίθηκαν ικανοποιητικές λόγω της προέλευσής τους και της αποτύπωσης λέξεων σχετικών με την Τεχνολογία Τροφίμων.

Εν τέλει, δημιουργήθηκε ένα σώμα κειμένων απαρτιζόμενο από 373 περιλήψεις στην Ιατρική, 223 στον Τουρισμό και 122 στην Τεχνολογία Τροφίμων. Όλα τα δεδομένα αποθηκεύτηκαν σε αρχείο .xlsb.

3.2. Διαχείριση δεδομένων

Το επόμενο βήμα ήταν η μείωση του θορύβου μέσα στο σώμα των κειμένων. Χρησιμοποιήθηκαν οι Μακροεντολές του Excel για τη δημιουργία του text handler (Triantafyllou et al., 2001) ο οποίος διαίρεσε τις γλωσσικές μονάδες και τα επιπλέον στοιχεία όπως ημερομηνίες και ακρωνύμια. Έπειτα, αφαιρέθηκαν οι τετριμμένες λέξεις και έγινε ανάλυση stemming στις εναπομένουσες. Πρέπει, βεβαίως, να επισημανθεί ότι η ομαδοποίηση των όρων έγινε με βαθμό ομοιότητας 66% και μηχανισμό εκτίμηση ομοιότητας τα 3-grams.

$$\text{Similarity}(W1,W2)=\text{CommonPositionTrigrams}\,(\text{Left}(W1, L), \text{Left}(W2, L))/ L$$
$$\text{όπου } L = (\text{Length}(W1) + \text{Length}(W2)) / 2 \quad , \quad L \epsilon N \tag{16}$$

Όροι εκπρόσωποι (πιο συχνή λέξη)	***Ομάδες όμοιων λέξεων***
ΤΟΥΡΙΣΜΟ	ΤΟΥΡΙΣΤΙΚΩΝ ΤΟΥΡΙΣΤΑ ΤΟΥΡΙΣΜΟΣ ΤΟΥΡΙΣΜΟΥ ΤΟΥΡΙΣΤΙΚΗ ΤΟΥΡΙΣΤΙΚΗΣ ΤΟΥΡΙΣΤΙΚΟΥ ΤΟΥΡΙΣΤΩΝ ΤΟΥΡΙΣΤΙΚΑ ΤΟΥΡΙΣΤΕΣ ΤΟΥΡΙΣΤΙΚΕΣ ΤΟΥΡΙΣΤΙΚΟΥΣ ΤΟΥΡΙΣΤΙΚΟΣ ΤΟΥΡΙΣΤΙΚΟΙ ΤΟΥΡΙΣΤΑΣ
ΝΟΣΟΚΟΜΕΙΟ	ΝΟΣΟΚΟΜΕΊΑ ΝΟΣΟΚΟΜΕΊΩΝ ΝΟΣΟΚΟΜΕΙΟΥ ΝΟΣΟΚΟΜΕΙΑΚΟ ΝΟΣΟΚΟΜΕΙΑΚΗ ΝΟΣΟΚΟΜΕΙΑΚΕΣ ΝΟΣΟΚΟΜΕΙΑΚΩΝ ΝΟΣΟΚΟΜΕΙΑΚΟΣ ΝΟΣΟΚΟΜΕΙΑΚΑ
ΑΣΘΕΝΕΙΣ	ΑΣΘΕΝΩΝ ΑΣΘΕΝΗ ΑΣΘΕΝΟΥΣ ΑΣΘΕΝΗΣ ΑΣΘΕΝΕΙΩΝ ΑΣΘΕΝΕΙΕΣ ΑΣΘΕΝΕΙΑ ΑΣΘΕΝΕΙΑΣ ΑΣΘΕΝΩΝΜΕ
ΓΥΝΑΙΚΑ	ΓΥΝΑΙΚΕΣ ΓΥΝΑΙΚΩΝ ΓΥΝΑΙΚΕΙΟΥ ΓΥΝΑΙΚΟΛΟΓΩΝ ΓΥΝΑΙΚΕΙΟΥΣ ΓΥΝΑΙΚΕΙΑ

Πίνακας 3.2-1. Παραδείγματα όμοιων λέξεων και των όρων - αντιπροσώπων τους.

3.3. Μέτρηση βαρύτητας, η DEVMAX.DF και η αναπαράσταση των όρων

Κατά την εφαρμογή της TF.IDF, η οποία επιλέχθηκε για αυτήν την έρευνα, παρατηρήθηκε ότι υψηλές θέσεις στη λίστα, καταλήφθηκαν από άσχετους όρους. Επακόλουθο ήταν η ανάγκη δημιουργίας μιας καινούργιας μετρικής για την παραγωγή μιας λίστας που θα βελτιώσει τα τελικά αποτελέσματα.

Η νέα μετρική ονομάζεται DEVMAX.DF. Βασίζεται στην προώθηση των όρων που εμφανίζονται σε μία ή περισσότερες κατηγορίες, αλλά όχι σε όλες. Αν ένας όρος εμφανίζεται πολλές φορές σε μια κατηγορία αλλά καθόλου ή ελάχιστα στις υπόλοιπες (μέγιστη απόκλιση), λαμβάνει καλύτερη θέση στη λίστα του μετρητή. Παράλληλα, η μετρική υποστηρίζει τους όρους με τη μεγαλύτερη συχνότητα με το λογάριθμο του DF, δηλαδή τον αριθμό των περιλήψεων που περιέχουν τον όρο F.

$$\text{DEVMAX. DF} = \sqrt{\frac{\sum_{i=1}^{c}(DF_i/N_i - \max)^2}{c * max^2}} * \log(DF), \quad (17)$$

$$\text{όπου } max = maximum_{i=1}^{c} DF_i/N_i$$

Στη νέα μετρική το DF_i είναι ο αριθμός των περιλήψεων που περιέχουν τον όρο F στην κλάση i. Το N_i είναι ο αριθμός των περιλήψεων στην κλάση i και c είναι ο αριθμός των κλάσεων.

Μετά την εφαρμογή της DEVMAX.DF, έγινε σύγκριση με την TF.IDF. Η λίστα των αποτελεσμάτων θεωρήθηκε βελτιωμένη καθώς έλλειπαν όροι που δεν αντιπροσωπεύουν τις κατηγορίες. Οι 10 πρώτες θέσεις των μετρικών μαζί με τη συχνότητα εμφάνισής τους στις τρεις κατηγορίες, παρουσιάζονται σε πίνακα (Πίνακας 3.3-2).

DEVMAX.DF				TF.IDF			
Όροι	*Ιατρική*	*Τουρισμός*	*Τρόφιμα*	*Όροι*	*Ιατρική*	*Τουρισμός*	*Τρόφιμα*
ΤΟΥΡΙΣΜΟ	0	187	0	ΤΟΥΡΙΣΜΟ	0	187	0
ΝΟΣΗΛΕΥΤΗΚΑΝ	129	0	0	ΑΣΘΕΝΩΝ	194	2	9
ΝΟΣΟΚΟΜΕΙΟ	101	0	0	ΝΟΣΗΛΕΥΤΗΚΑΝ	129	0	0
ΑΣΘΕΝΩΝ	194	2	9	ΥΓΕΙΑΣ	147	5	14
ΦΡΟΝΤΙΔΑ	70	0	0	ΠΑΙΔΙΑ	49	1	4
ΓΥΝΑΙΚΕΣ	68	1	0	ΑΝΑΠΤΥΣΣΕΙ	66	112	48
ΚΛΙΝΙΚΗ	85	1	2	ΠΟΙΟΤΗΤΑΣ	85	39	28
ΤΡΟΦΙΜΩΝ	2	1	56	ΜΕΘΟΔΟΥΣ	204	34	56
ΘΕΡΑΠΕΙΑΣ	104	3	4	ΑΝΑΓΚΕΣ	151	88	53
ΑΝΑΣΚΟΠΗΣΗ	98	8	1	ΕΚΠΑΙΔΕΥΤΙΚΩΝ	88	14	2

Πίνακας 3.3-2. Οι λίστες με τους 10 πρώτους όρους των μετρικών βαρύτητας και η εμφάνισή τους στις 3 κατηγορίες.

Η εμφάνιση των όρων στις περιλήψεις εκφράστηκε με δύο τρόπους:

- την πραγματική συχνότητα εμφάνισης τους (tf)
- τη δυαδική αναπαράσταση παρουσίας τους ή όχι (bin).

Ως εκ τούτου, χρησιμοποιήθηκαν τέσσερις πειραματικές μέθοδοι, οι TF.IDF-bin, TF.IDF-tf, DEVMAX.DF-bin και DEVMAX.DF-tf. Η δειγματοληψία έγινε στους πρώτους 10, 15, 20, 25, 50, 75, 100, 150, 200, 300, 500 και 750 όρους (W).

3.4. Κατηγοριοποίηση

Αρχικά, η κατηγοριοποίηση έγινε με όλους τους αλγορίθμους που είναι διαθέσιμοι στο WEKA, έκδοση 3.7.12. Ωστόσο, πολλοί από αυτούς αποκλείστηκαν λόγω της πολύ χαμηλής απόδοσής τους ανεξαρτήτως διανυσμάτων.

Οι εναπομείναντες αλγόριθμοι ήταν:

- Δύο Bayesian: NaïveBayes και NaïveBayesMultinomial,
- Τρείς Function: MultilayerPerceptron, SimpleLogistic, και SMO(SVM),
- Δύο Lazy: IBk και Kstar,
- Δύο Metalearning: ClassificationViaRegression και LogitBoost,
- Τρεις Rule classifiers: DecisionTable, JRip, και PART,
- Δύο Tree: LMT και RandomForest.

Επιλέχθηκε η τεχνική επικύρωσης k-fold cross validation όπου το δείγμα χωρίζεται τυχαία σε k ίσα μέρη από τα οποία τα k-1 χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου ενώ το εναπομένον υποστηρίζει την αξιολόγηση. Αυτή η

διαδικασία συμβαίνει k φορές (Kohavi, 1995). Επιλέχθηκε η 10-fold cross validation, η οποία είναι προεπιλεγμένη στο WEKA. Η αξιολόγηση έγινε με βάση την Ακρίβεια, την Ανάκληση και το F-score.

3.5. Συσταδοποίηση

Παρόμοια με την κατηγοριοποίηση, και κατά αυτήν την εργασία δοκιμάστηκαν όλοι οι συσταδοποιητές του προγράμματος, όμως κάποιοι αποκλείστηκαν λόγω των αποτελεσμάτων τους. Τα αποτελέσματα που παρουσιάζονται προέρχονται από τους συσταδοποιητές:

- Canopy,
- Cascade Simpe KMeans (CascadeSKM)
- Expectation Maximization (EM)
- Self-organizing maps (SOM)
- XMeans.

Ωστόσο, η μεθοδολογία που ακολουθήθηκε στη συσταδοποίηση διαφέρει. Ο τρόπος συσταδοποίησης βασίστηκε στο classes to clusters evaluation, όπου τα αποτελέσματα της συσταδοποίησης παραβάλλονται με την ήδη υπάρχουσα κατηγοριοποίηση. Κατά συνέπεια αποδίδεται ένα ποσοστό λάθος συσταδοποιημένων παραδειγμάτων (Incorrectly Clustered Instances - ICI), το οποίο αποτελεί τη βάση αξιολόγησης των συσταδοποιητών.

Αρχικά, ορίστηκε το κατώτατο όριο των τριών συστάδων για όλους τους αλγορίθμους, καθώς το δείγμα αποτελείται από τόσες κατηγορίες. Έπειτα, οι συστάδες αυξομειώνονταν μέχρι να βρεθεί ο αριθμός που θα φέρει το μικρότερο ποσοστό λάθος κατηγοριοποιημένων παραδειγμάτων. Επισημαίνεται ότι σε αυτή την περίπτωση, το πρόγραμμα παρουσίαζε ως λάθη όσα παραδείγματα υπήρχαν σε άλλες συστάδες, πέραν των τριών που αφορούσαν τις μεγάλες κατηγορίες. Καθίστατο, λοιπόν, απαραίτητη η μέτρηση τους ως σωστά και η αφαίρεσή τους από τα ποσοστά λάθους.

Πιο συγκεκριμένα, σύμφωνα με το παράδειγμα που ακολουθεί στα 718 παραδείγματα (Πίνακας 3.5-3), δημιουργήθηκαν 4 συστάδες εκ των οποίων η 0 δεν σχετίζεται με καμία κατηγορία κατά το WEKA. Ωστόσο, είναι εμφανές, ότι η συγκεκριμένη συστάδα περιέχει 140 παραδείγματα που ανήκουν στην Ιατρική. Ως εκ τούτου, πρέπει από τα 188 (26,18%) λάθος συσταδοποιημένα παραδείγματα να αφαιρεθούν τα 140, αφήνοντας 48 (5,84%). Το τελικό ICI διαμορφώνεται στα 5,84%.

Συστάδες				*Κατηγορίες*
0	1	2	3	
140	220	13	0	Ιατρική
3	0	9	110	Τεχν.Τροφίμων
0	10	200	13	Τουρισμός

Συστάδα 0 → No class

Συστάδα 1 → Ιατρική

Συστάδα 2 → Τουρισμός

Συστάδα 3 → Τεχν.Τροφίμων ICI: 188 26,18% → 48 5,84%

Πίνακας 3.5-3. Παράδειγμα αποτελέσματος συσταδοποίησης όπως παρουσιάζεται στο WEKA.

4. Αποτελέσματα

4.1. Κατηγοριοποίηση

Οι Πίνακας 4.1-4 και Πίνακας 4.1-5 παρουσιάζουν αναλυτικά τα αποτελέσματα της κατηγοριοποίησης. Αν και, γενικά, η διαδικασία κύλησε ομαλά, παρουσιάστηκε πρόβλημα κατά τη μέτρηση με τη χρήση του Multilayer Perceptron (MLP) στα 500W και 750W και στις τέσσερις πειραματικές μεθόδους. Πρέπει να σημειωθεί ότι η πολυπλοκότητα των μοντέλων τύπου τεχνητών νευρωνικών δικτύων, αυξάνει δραματικά το χρόνο που απαιτούν για την ολοκλήρωση μιας εργασίας. Μετά τις 3,5 ώρες, λοιπόν, το WEKA έπαυε τη λειτουργία του κατηγοριοποιητή χωρίς να εμφανίσει αποτελέσματα.

	Διάνυσμα / *Κατηγοριοποιητής*	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	NaïveBayes	89,90	92,49	92,84	92,95	93,55	94,05	94,35	95,10	95,45	95,25	95,60	95,75
	NBMultinomial	87,03	89,25	91,80	94,20	95,05	95,85	96,00	96,10	96,15	95,75	95,85	96,30
	MLP	89,47	92,89	92,30	92,35	93,80	94,20	94,50	96,05	95,45	96,55	Failed	Failed
	SimpleLogistic	89,96	92,89	91,80	94,30	96,10	96,40	96,95	97,10	96,40	96,90	96,50	96,00
	SMO	89,03	92,94	91,80	93,30	95,30	96,40	96,00	96,20	96,10	97,20	97,10	96,70
	IBk	89,90	92,64	92,65	93,50	92,40	91,85	92,30	90,50	85,79	82,08	73,04	71,12
	Kstar	90,16	92,89	92,54	92,20	92,35	91,85	92,15	90,60	87,24	84,20	76,43	73,21
	Class.ViaRegress.	86,15	86,39	88,69	90,45	93,50	94,20	94,60	94,45	93,65	95,15	95,15	95,15
	LogitBoost	87,23	90,69	91,80	94,30	94,60	96,15	95,45	96,25	96,25	96,00	96,05	96,05
	DecisionTable	86,83	88,99	90,75	91,60	91,45	91,00	91,75	90,95	90,95	91,95	91,65	91,40
	JRip	86,49	92,39	90,85	91,95	92,20	93,10	94,10	93,65	93,25	93,55	92,95	93,50
	PART	89,75	91,99	92,69	92,50	92,25	93,50	94,85	94,05	93,95	93,20	93,60	94,30
	LMT	89,96	92,89	93,15	94,30	96,10	96,40	96,80	96,90	96,20	96,90	96,20	96,00
	RandomForest	90,00	92,84	93,10	92,95	93,75	94,60	95,80	96,40	97,10	97,50	96,85	97,15
TF	NaïveBayes	79,94	87,65	87,23	90,54	90,39	91,09	91,35	91,89	92,80	94,15	94,75	94,95
	NBMultinomial	87,23	89,44	92,55	94,80	95,65	95,75	95,90	96,45	96,30	96,45	96,05	97,20
	MLP	88,44	91,65	91,70	93,00	93,55	93,70	92,55	93,00	92,45	91,25	Failed	Failed
	SimpleLogistic	87,20	92,69	92,80	95,15	95,40	96,20	96,00	96,00	96,00	94,55	95,40	94,60
	SMO	80,79	83,85	89,19	91,25	95,05	95,15	94,85	93,40	94,70	94,80	95,55	95,05
	IBk	89,49	91,84	92,09	93,30	88,05	86,95	85,85	86,25	85,75	80,04	77,48	73,24
	Kstar	90,16	92,60	91,60	92,30	90,63	89,66	88,91	87,74	83,83	79,94	74,35	72,00
	Class.ViaRegress.	87,28	86,98	88,54	89,60	93,05	92,90	93,15	93,25	93,55	93,75	93,80	94,00
	LogitBoost	87,23	90,69	91,80	94,20	94,85	95,60	95,60	96,25	96,25	95,70	95,30	95,30
	DecisionTable	86,83	88,99	90,75	91,15	91,25	90,90	91,75	90,95	90,95	91,95	91,65	91,40
	JRip	86,74	92,34	90,70	90,80	92,15	93,30	92,95	93,65	93,60	94,30	93,65	93,35
	PART	89,39	91,94	92,94	92,95	93,35	93,25	93,75	94,20	94,20	93,75	94,20	92,90
	LMT	89,60	92,29	93,20	95,15	95,40	96,20	95,80	95,10	96,00	94,55	95,10	94,30
	RandomForest	89,95	92,59	92,29	92,95	93,55	94,30	96,40	96,55	96,40	97,55	97,60	96,55

Πίνακας 4.1-4. F-score (%) με όρους από τη DEVMAX.DF.

	Διάνυσμα / *Κατηγοριοποιητής*	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	NaïveBayes	83,91	83,50	84,55	86,85	92,25	92,95	93,25	93,05	93,30	94,80	93,30	95,75
	NBMultinomial	77,28	82,25	85,45	88,75	93,80	94,90	94,65	94,75	93,20	95,50	95,10	96,30
	MLPerceptron	81,89	82,60	83,90	87,50	92,90	95,10	95,05	95,15	95,55	96,25	Failed	Failed
	SimpleLogistic	80,35	83,15	86,10	87,65	93,50	94,85	95,55	95,85	96,70	95,70	96,40	96,00
	SMO	84,68	83,50	86,00	87,50	92,20	93,30	93,60	94,60	95,70	95,85	95,75	96,70
	IBk	81,59	80,60	80,55	85,60	86,00	86,50	87,14	83,18	80,21	79,32	67,81	71,12
	Kstar	81,70	81,00	82,75	86,35	87,00	88,45	87,69	84,48	81,96	80,70	70,38	73,21
	Class.ViaRegress.	81,65	84,55	86,25	86,95	91,85	93,70	93,80	93,35	93,55	94,00	93,65	95,15
	LogitBoost	81,65	82,40	84,80	88,25	92,40	93,95	94,50	94,65	94,35	96,00	95,75	96,05
	DecisionTable	82,31	81,50	83,25	81,60	88,95	92,45	92,00	92,00	91,70	92,00	92,10	91,40
	JRip	79,50	81,60	83,65	83,30	90,15	91,30	93,15	92,00	92,70	90,40	91,95	93,50
	PART	82,22	81,90	84,15	86,70	90,00	91,95	92,05	92,25	92,95	92,60	93,10	94,30
	LMT	80,75	82,80	86,25	87,65	93,50	94,85	96,00	95,85	96,50	95,70	96,40	96,00
	RandomForest	82,23	82,35	86,10	89,15	93,60	95,80	96,70	96,25	96,70	97,40	96,55	97,15
TF	NaïveBayes	74,00	75,87	77,69	80,22	85,73	87,94	89,24	90,09	90,95	92,75	93,00	92,95
	NBMultinomial	81,29	83,29	86,03	87,14	92,50	94,75	94,50	95,20	95,75	97,25	96,70	96,55
	MLP	80,80	81,80	84,05	87,85	91,60	94,80	92,85	93,35	91,70	84,80	Failed	Failed
	SimpleLogistic	82,10	84,50	86,90	87,90	93,70	94,40	95,15	94,20	94,70	95,00	95,30	95,00
	SMO	76,88	78,86	80,95	83,77	90,15	93,05	92,55	92,95	93,35	94,30	92,90	94,05
	IBk	75,70	75,85	76,20	80,00	79,40	82,50	79,64	78,24	75,81	75,93	71,97	66,04
	Kstar	79,60	77,60	79,75	80,35	80,09	80,67	77,11	73,43	72,18	70,14	60,52	57,89
	Class.ViaRegress.	81,30	84,55	86,05	87,15	90,05	92,75	92,95	91,60	92,25	92,25	92,25	92,25
	LogitBoost	80,75	83,80	85,80	87,85	92,60	94,65	93,95	94,25	94,35	96,00	95,70	95,30
	DecisionTable	82,02	82,95	81,70	81,85	89,50	92,45	91,85	91,45	91,45	91,80	91,85	92,00
	JRip	80,75	81,05	81,70	83,15	90,25	91,95	92,05	92,65	91,95	91,35	91,55	91,55
	PART	80,90	81,85	83,40	83,85	90,85	92,25	91,65	92,15	92,05	91,50	91,40	90,75
	LMT	82,10	84,20	86,90	87,90	93,70	94,70	95,00	94,30	94,70	95,00	95,30	95,00
	RandomForest	80,99	85,45	87,60	89,70	93,20	95,40	96,80	96,15	96,55	96,25	96,95	97,40

Πίνακας 4.1-5. F-score(%) με όρους από τη TF.IDF.

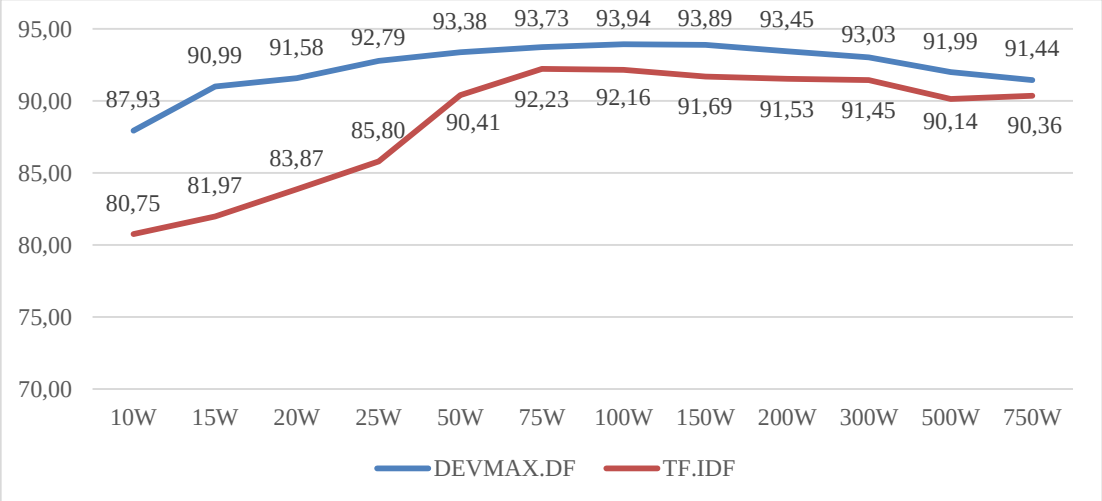
Εντούτοις, ο RandomForest απέφερε τα καλύτερα αποτελέσματα, συγκριτικά με όλους τους κατηγοριοποιητές, σε όλες τις πειραματικές μεθόδους (Πίνακας 4.1-6). Απέδωσε το υψηλότερο F-score με τη μέθοδο DEVMAX.DF-tf, F1=97,60%, ενώ ακολουθούν οι μέθοδοι DEVMAX.DF-bin με F1=97,50%, TF.IDF-bin και TF.IDF-tf με F1=97,40%. Έπεται ο NaiveBayes Multinomial (NBMultinomial) με F1=97,25%

στην TF.IDF-tf και F1=97,20% στην DEVMAX.DF-tf. Ο SMO και ο SimpleLogistic απέδωσαν F1=97,20% και FI=97,10% έκαστος στην DEVMAX.DF-bin.

ΚΑΤΗΓΟΡΙΟ-ΠΟΙΗΤΗΣ	*ΜΕΘΟΔΟΣ*	*ΔΙΑΝΥΣΜΑ*	*F-SCORE*	*ΑΚΡΙΒΕΙΑ*	*ΑΝΑΚΛΗΣΗ*
RandomForest	DEVMAX.DF-tf	500W	97,60	97,60	97,60
RandomForest	DEVMAX.DF-bin	300W	97,50	97,50	97,50
RandomForest	TF.IDF-bin	300W	97,40	97,40	97,40
RandomForest	TF.IDF-tf	750W	97,40	97,40	97,40
NBMultinomial	TF.IDF-tf	300W	97,25	97,30	97,20
NBMultinomial	DEVMAX.DF-tf	750W	97,20	97,20	97,20
SMO	DEVMAX.DF-bin	300W	97,20	97,20	97,20
SimpleLogistic	DEVMAX.DF-bin	150W	97,10	97,10	97,10

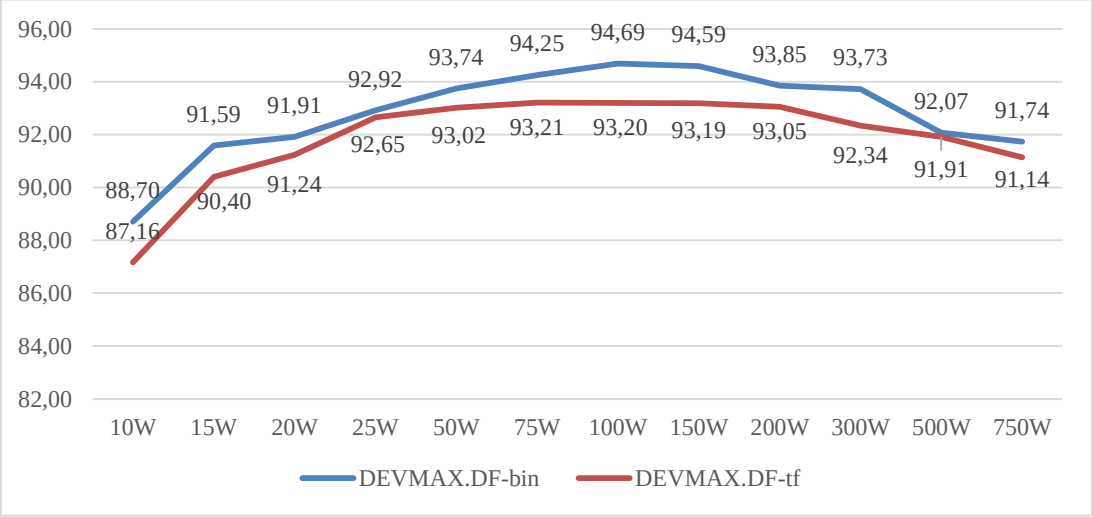
Πίνακας 4.1-6. Αποτελέσματα (%) των αποδοτικότερων κατηγοριοποιητών με τα αντίστοιχα διανύσματα.

Συγκρίνοντας τις δύο μετρικές βαρύτητας με τη χρήση όλων των F-scores, συνάγεται ότι η DEVMAX.DF είχε καλύτερες επιδόσεις από την TF.IDF. Το Διάγραμμα 4.1-1 δείχνει τη διαφορά τους, ιδιαίτερα στα μικρά διανύσματα όπου η DEVMAX.DF κατάφερε να εντοπίσει τις πιο εύστοχες λέξεις, ενώ η TF.IDF ανέβασε την απόδοσή της από τους 25 όρους και πάνω.

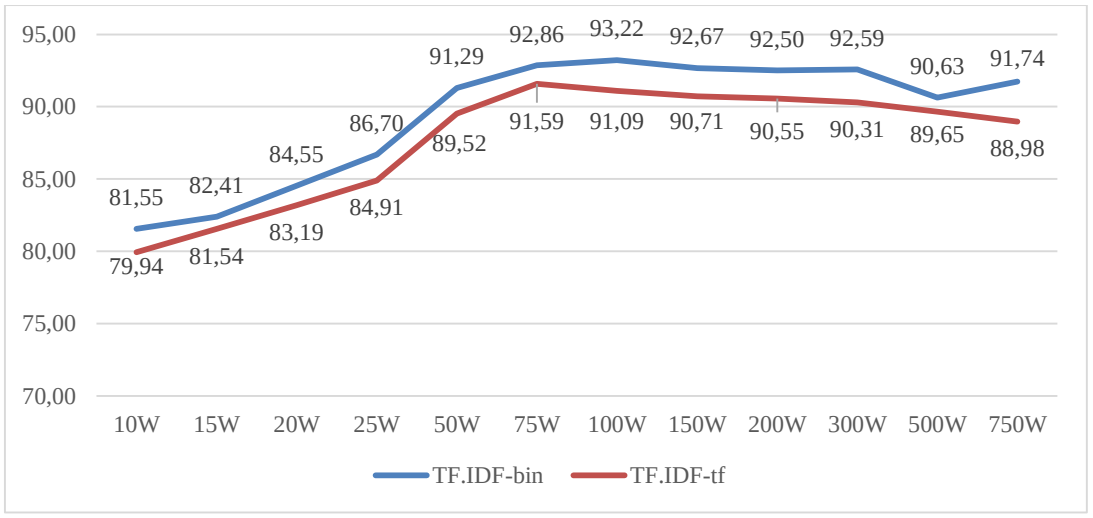


Διάγραμμα 4.1-1. Μέσο F-score (%) απόδοσης όλων των κατηγοριοποιητών.

Παράλληλα, τα Διάγραμμα 4.1-2 και Διάγραμμα 4.1-3 παρουσιάζουν την απόδοση της δυαδικής εμφάνισης (μπλε) έναντι στην συχνότητα εμφάνισης (κόκκινο). Η χρήση της δυαδικής εμφάνισης ωθεί σε ελαφρώς αυξημένα αποτελέσματα και στις δύο μετρικές.



Διάγραμμα 4.1-2. Μέση τιμή F-score (%) όλων των κατηγοριοποιητών για την δυαδική εμφάνιση (bin) και τη συχνότητα εμφάνισης (tf) για την DEVMAX.DF.



Διάγραμμα 4.1-3. Μέση τιμή F-score (%) όλων των κατηγοριοποιητών για την δυαδική εμφάνιση (bin) και τη συχνότητα εμφάνισης (tf) για την TF.IDF.

4.2. Συσταδοποίηση

Οι Πίνακας 4.2-7 και Πίνακας 4.2-9 παρουσιάζουν αναλυτικά τα αποτελέσματα της συσταδοποίησης. Παρόμοια, ο SOM, ο οποίος βασίζεται στα τεχνητά νευρωνικά δίκτυα, εμφάνισε πρόβλημα στην εξαγωγή αποτελεσμάτων. Η προσπάθεια ρύθμισης των συστάδων, από τέσσερις και πάνω, αύξανε τον απαιτούμενο χρόνο. Κατά συνέπεια, μετά τις 3,5 ώρες το πρόγραμμα τερμάτιζε τη λειτουργία του αλγορίθμου χωρίς να παρουσιάζει αποτελέσματα. Εξαιτίας αυτού, ο SOM είναι ο μόνος αλγόριθμος που δεν απέδωσε αποτελέσματα με πάνω από τρεις συστάδες στους 750 όρους.

Κατά τη συσταδοποίηση μετρήθηκε και ο αριθμός των συστάδων που αποδίδουν τα πιο αμιγή αποτελέσματα, όπως φαίνονται στους Πίνακας 4.2-8 και Πίνακας 4.2-10.

	Διάνυσμα / *Συσταδοποιητής*	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	Canopy	20,75	14,76	11,42	12,26	24,10	25,07	30,08	40,11	47,77	48,75	46,80	48,47
	CascadeSKM	14,21	20,47	11,14	11,42	11,00	6,13	6,41	5,71	5,99	6,83	6,83	13,51
	EM	21,87	17,83	14,76	18,80	12,54	10,17	8,50	7,80	4,88	7,10	6,13	5,57
	SOM	14,21	11,42	11,14	10,86	10,03	7,24	6,55	7,24	8,08	8,08	4,46	19,92
	XMEANS	14,21	21,45	18,80	12,81	8,22	17,83	14,35	14,62	20,33	20,75	20,61	21,59
TF	Canopy	44,29	44,29	39,97	36,07	42,06	34,96	35,10	46,10	48,19	47,08	44,71	47,35
	CascadeSKM	27,16	29,53	27,72	23,12	24,65	24,23	27,72	31,06	33,43	34,40	29,94	33,15
	EM	30,78	30,78	17,13	15,18	12,12	11,70	12,81	13,79	12,54	16,30	17,83	19,08
	SOM	27,16	29,53	31,76	27,58	30,22	25,91	29,11	31,34	26,88	21,03	20,89	47,21
	XMEANS	19,92	24,65	28,27	22,84	25,35	28,13	22,07	29,09	29,11	46,38	31,75	33,01

Πίνακας 4.2-7. Λάθος συσταδοποιημένα παραδείγματα (ICI)(%) με την DEVMAX.DF.

	Διάνυσμα / *Συσταδοποιητής*	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	Canopy	6	6	5	6	7	6	6	5	4	4	4	4
	CascadeSKM	7	7	6	6	5	5	5	5	5	5	6	4
	EM	4	4	3	3	4	5	5	6	5	4	4	5
	SOM	8	7	5	5	5	4	4	4	5	5	6	3
	XMEANS	6	5	3	3	6	6	4	4	4	6	6	3
TF	Canopy	12	12	11	9	9	8	7	6	8	7	5	5
	CascadeSKM	15	14	12	11	11	10	9	9	8	6	5	5
	EM	12	11	11	11	10	9	6	6	6	6	5	4
	SOM	12	11	11	11	9	9	8	7	7	7	7	3
	XMEANS	18	17	16	14	12	10	8	8	8	7	3	6

Πίνακας 4.2-8. Αριθμός συστάδων με την DEVMAX.DF.

	Διάνυσμα Συσταδοποιητής	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	Canopy	30,92	28,13	28,83	30,22	29,11	26,04	32,73	35,52	42,48	38,44	45,27	48,19
	CascadeSKM	24,37	23,12	27,30	22,42	12,12	10,45	10,86	12,40	22,01	21,59	25,35	24,23
	EM	32,73	28,69	28,83	19,22	12,95	12,81	10,59	10,86	11,28	5,57	12,40	16,57
	SOM	25,91	28,55	25,35	21,59	14,35	10,03	9,75	9,75	9,47	9,61	9,61	21,31
	XMEANS	18,11	25,77	23,82	23,26	17,69	15,74	9,33	21,87	22,98	22,14	24,51	25,07
TF	Canopy	47,35	48,19	48,05	49,44	47,35	47,35	48,89	48,47	48,61	48,33	48,33	48,33
	CascadeSKM	27,58	30,36	30,64	34,26	37,88	32,17	23,68	32,03	31,34	35,24	37,60	39,83
	EM	32,73	26,74	32,87	33,43	20,19	15,60	13,51	13,93	15,18	15,04	19,22	22,70
	SOM	30,02	31,62	24,37	32,31	26,04	27,16	21,03	25,77	28,13	23,82	24,65	45,68
	XMEANS	32,31	35,65	33,84	34,54	31,89	33,57	37,74	36,21	35,38	38,02	39,14	38,86

Πίνακας 4.2-9. Λάθος συσταδοποιημένα παραδείγματα (ICI) (%) με την TF.IDF.

	Διάνυσμα Συσταδοποιητής	*10W*	*15W*	*20W*	*25W*	*50W*	*75W*	*100W*	*150W*	*200W*	*300W*	*500W*	*750W*
BIN	Canopy	14	14	14	12	10	9	9	8	6	3	4	3
	CascadeSKM	9	9	9	9	8	8	7	6	5	3	3	3
	EM	4	5	5	3	9	8	8	7	6	5	5	4
	SOM	11	11	10	8	8	8	8	7	7	6	5	3
	XMEANS	9	9	9	9	8	8	6	6	6	6	6	6
TF	Canopy	4	4	4	4	4	3	3	3	3	3	3	3
	CascadeSKM	12	12	12	10	9	9	8	7	7	7	5	4
	EM	8	8	7	5	5	5	5	5	7	7	4	4
	SOM	14	14	13	13	12	12	11	9	8	7	6	3
	XMEANS	9	9	8	8	6	6	3	4	4	4	3	3

Πίνακας 4.2-10. Αριθμός συστάδων με την TF.IDF.

Το καλύτερο αποτέλεσμα απέδωσε ο SOM με ICI=4,46% με τη μέθοδο DEVMAX.DF-bin στις 6 συστάδες. Στην δεύτερη, τρίτη και τέταρτη θέση ακολουθεί ο EM με ICI=4,88% και 5,57% με τη μέθοδο DEVMAX.DF-bin και ICI=5,57% στην TF.IDF-bin στις 5 συστάδες. Έπεται ο CascadeSKM με ICI=5,71% και 5,99% με τη DEVMAX.DF-bin στις 5 συστάδες. Τα αναλυτικά αποτελέσματα των συσταδοποιητών με ICI<10% παρουσιάζονται αναλυτικά σε πίνακα (Πίνακας 4.2-11).

Πίνακας 4.2-11. Αποτελέσματα των καλύτερων συσταδοποιητών.

Συσταδοποιητής	**Μέθοδος**	**Διάνυσμα**	**Ποσοστό λάθους (ICI)(%)**	**Συστάδες**

SOM	DEVMAX.DF-bin	500W	4,46	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				122 88 0 5 158 0 | Iatrikh
				  0 0 0 1 21 201 | Tourismos
				  1 0 61 56 1 3 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- No class
				Cluster 2 <-- Trofima
				Cluster 3 <-- No class
				Cluster 4 <-- Iatrikh
				Cluster 5 <-- Tourismos

EM	DEVMAX.DF-bin	200W	4,88	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  7 125 130 100 11 | Iatrikh
				  1 0 2 3 217 | Tourismos
				111 0 0 1 10 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- Iatrikh
				Cluster 3 <-- No class
				Cluster 4 <-- Tourismos

EM	DEVMAX.DF-bin	750W	5,57	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				 96 3 9 119 146 | Iatrikh
				  3 1 218 0 1 | Tourismos
				 16 99 5 1 1 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- Trofima
				Cluster 2 <-- Tourismos
				Cluster 3 <-- No class
				Cluster 4 <-- Iatrikh

EM	TF.IDF-bin	300W	5,57	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				120 3 2 146 102 | Iatrikh
				  8 0 212 1 2 | Tourismos
				 11 98 12 1 0 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- Trofima
				Cluster 2 <-- Tourismos
				Cluster 3 <-- Iatrikh
				Cluster 4 <-- No class

CascadeSKM	DEVMAX.DF-bin	150W	5,71	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  0 116 173 84 0 | Iatrikh
				  0 0 19 0 204 | Tourismos
				100 1 18 0 3 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- Iatrikh
				Cluster 3 <-- No class
				Cluster 4 <-- Tourismos

CascadeSKM	DEVMAX.DF-bin	200W	5,99	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  0 168 90 113 2 | Iatrikh
				204 19 0 0 0 | Tourismos
				  4 17 0 1 100 | Trofima
				
				Cluster 0 <-- Tourismos
				Cluster 1 <-- Iatrikh
				Cluster 2 <-- No class
				Cluster 3 <-- No class
				Cluster 4 <-- Trofima

CascadeSKM	DEVMAX.DF-bin	75W	6,13	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  7 116 160 90 0 | Iatrikh
				  1 0 19 0 203 | Tourismos
				105 0 14 0 3 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- Iatrikh
				Cluster 3 <-- No class
				Cluster 4 <-- Tourismos

EM	DEVMAX.DF-bin	500W	6,13	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 <-- assigned to cluster
				209 3 3 158 | Iatrikh
				  0 0 211 12 | Tourismos
				  0 96 4 22 | Trofima
				
				Cluster 0 <-- Iatrikh
				Cluster 1 <-- Trofima
				Cluster 2 <-- Tourismos
				Cluster 3 <-- No class

CascadeSKM	DEVMAX.DF-bin	100W	6,41	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  5 92 0 159 117 | Iatrikh
				  2 0 198 23 0 | Tourismos
				106 0 2 13 1 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- Tourismos
				Cluster 3 <-- Iatrikh
				Cluster 4 <-- No class

SOM	DEVMAX.DF-bin	100W	6,55	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 <-- assigned to cluster
				  0 3 164 206 | Iatrikh
				197 1 25 0 | Tourismos
				  1 104 17 0 | Trofima
				
				Cluster 0 <-- Tourismos
				Cluster 1 <-- Trofima
				Cluster 2 <-- No class
				Cluster 3 <-- Iatrikh

CascadeSKM	DEVMAX.DF-bin	300W	6,83	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  0 123 163 87 0 | Iatrikh
				  0 0 25 0 198 | Tourismos
				 98 1 21 0 2 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- Iatrikh
				Cluster 3 <-- No class
				Cluster 4 <-- Tourismos

CascadeSKM	DEVMAX.DF-bin	500W	6,83	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				  0 1 92 166 0 114 | Iatrikh
				  0 0 0 14 209 0 | Tourismos
				 87 0 0 27 7 1 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- No class
				Cluster 2 <-- No class
				Cluster 3 <-- Iatrikh
				Cluster 4 <-- Tourismos
				Cluster 5 <-- No class

EM	DEVMAX.DF-bin	300W	7,10	Class attribute: CLASS Classes to Clusters: 0 1 2 3 <-- assigned to cluster 26 5 198 144 | Iatrikh 218 1 2 2 | Tourismos 14 107 0 1 | Trofima Cluster 0 <-- Tourismos Cluster 1 <-- Trofima Cluster 2 <-- Iatrikh Cluster 3 <-- No class

SOM	DEVMAX.DF-bin	150W	7,24	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 <-- assigned to cluster
				  0 168 0 205 | Iatrikh
				195 28 0 0 | Tourismos
				  2 21 98 1 | Trofima
				
				Cluster 0 <-- Tourismos
				Cluster 1 <-- No class
				Cluster 2 <-- Trofima
				Cluster 3 <-- Iatrikh

SOM	DEVMAX.DF-bin	75W	7,24	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 <-- assigned to cluster
				209 153 11 0 | Iatrikh
				  0 23 1 199 | Tourismos
				  0 15 105 2 | Trofima
				
				Cluster 0 <-- Iatrikh
				Cluster 1 <-- No class
				Cluster 2 <-- Trofima
				Cluster 3 <-- Tourismos

EM	DEVMAX.DF-bin	150W	7,80	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				106 7 84 22 143 11 | Iatrikh
				  7 2 0 213 0 1 | Tourismos
				  5 105 0 11 0 1 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- Trofima
				Cluster 2 <-- No class
				Cluster 3 <-- Tourismos
				Cluster 4 <-- Iatrikh
				Cluster 5 <-- No class

SOM	DEVMAX.DF-bin	200W	8,08	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  0 163 0 90 120 | Iatrikh
				194 29 0 0 0 | Tourismos
				  2 26 93 0 1 | Trofima
				
				Cluster 0 <-- Tourismos
				Cluster 1 <-- Iatrikh
				Cluster 2 <-- Trofima
				Cluster 3 <-- No class
				Cluster 4 <-- No class

SOM	DEVMAX.DF-bin	300W	8,08	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				  0 159 0 95 119 | Iatrikh
				195 28 0 0 0 | Tourismos
				  3 26 92 0 1 | Trofima
				
				Cluster 0 <-- Tourismos
				Cluster 1 <-- Iatrikh
				Cluster 2 <-- Trofima
				Cluster 3 <-- No class
				Cluster 4 <-- No class

XMEANS	DEVMAX.DF-bin	50W	8,22	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				129 27 12 69 79 57 | Iatrikh
				  3 6 214 0 0 0 | Tourismos
				  5 111 5 0 0 1 | Trofima
				
				Cluster 0 <-- Iatrikh
				Cluster 1 <-- Trofima
				Cluster 2 <-- Tourismos
				Cluster 3 <-- No class
				Cluster 4 <-- No class
				Cluster 5 <-- No class

EM	DEVMAX.DF-bin	100W	8,50	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 <-- assigned to cluster
				 19 22 85 19 228 | Iatrikh
				  0 7 0 212 4 | Tourismos
				  1 113 0 8 0 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- Trofima
				Cluster 2 <-- No class
				Cluster 3 <-- Tourismos
				Cluster 4 <-- Iatrikh

XMEANS	TF.IDF-bin	100W	9,33	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				  6 4 100 106 118 39 | Iatrikh
				  3 187 1 4 22 6 | Tourismos
				101 4 0 0 14 3 | Trofima
				
				Cluster 0 <-- Trofima
				Cluster 1 <-- Tourismos
				Cluster 2 <-- No class
				Cluster 3 <-- No class
				Cluster 4 <-- Iatrikh
				Cluster 5 <-- No class

SOM	TF.IDF-bin	200W	9,47	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 6 <-- assigned to cluster
				135 91 1 4 4 10 128 | Iatrikh
				  0 0 0 32 81 106 4 | Tourismos
				  1 1 77 22 0 6 15 | Trofima
				
				Cluster 0 <-- Iatrikh
				Cluster 1 <-- No class
				Cluster 2 <-- Trofima
				Cluster 3 <-- No class
				Cluster 4 <-- No class
				Cluster 5 <-- Tourismos
				Cluster 6 <-- No class

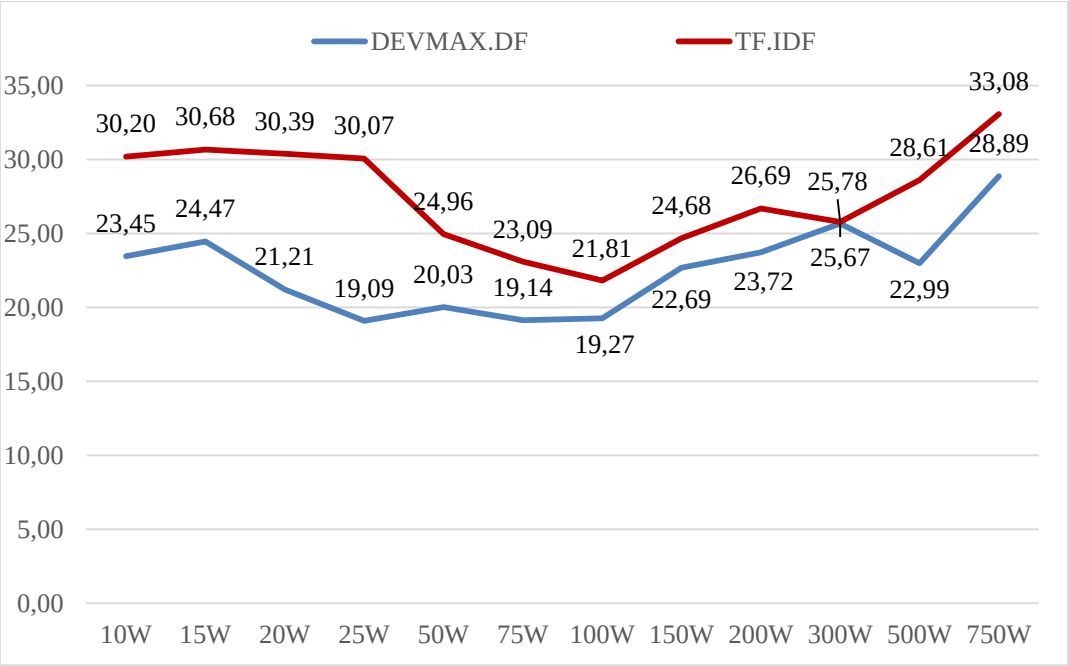
SOM	TF.IDF-bin	300W	9,61	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 <-- assigned to cluster
				135 93 2 5 2 136 | Iatrikh
				  0 0 0 62 142 19 | Tourismos
				  1 0 81 10 1 29 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- No class
				Cluster 2 <-- Trofima
				Cluster 3 <-- No class
				Cluster 4 <-- Tourismos
				Cluster 5 <-- Iatrikh

SOM	TF.IDF-bin	500W	9,61	Class attribute: CLASS Classes to Clusters: 0 1 2 3 4 <-- assigned to cluster 136 5 7 1 224 | Iatrikh 10 158 55 0 0 | Tourismos 32 4 9 76 1 | Trofima Cluster 0 <-- No class Cluster 1 <-- Tourismos Cluster 2 <-- No class Cluster 3 <-- Trofima Cluster 4 <-- Iatrikh

SOM	TF.IDF-bin	100W	9,75	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 6 7 <-- assigned to cluster
				114 94 61 2 1 3 2 96 | Iatrikh
				  1 0 3 0 16 62 123 18 | Tourismos
				  0 1 7 72 26 0 1 15 | Trofima
				
				Cluster 0 <-- Iatrikh
				Cluster 1 <-- No class
				Cluster 2 <-- No class
				Cluster 3 <-- Trofima
				Cluster 4 <-- No class
				Cluster 5 <-- No class
				Cluster 6 <-- Tourismos
				Cluster 7 <-- No class

SOM	TF.IDF-bin	150W	9,75	Class attribute: CLASS
				Classes to Clusters:
				
				  0 1 2 3 4 5 6 <-- assigned to cluster
				128 10 2 5 1 97 130 | Iatrikh
				  5 103 82 32 0 0 1 | Tourismos
				 11 6 0 27 76 1 1 | Trofima
				
				Cluster 0 <-- No class
				Cluster 1 <-- Tourismos
				Cluster 2 <-- No class
				Cluster 3 <-- No class
				Cluster 4 <-- Trofima
				Cluster 5 <-- No class
				Cluster 6 <-- Iatrikh

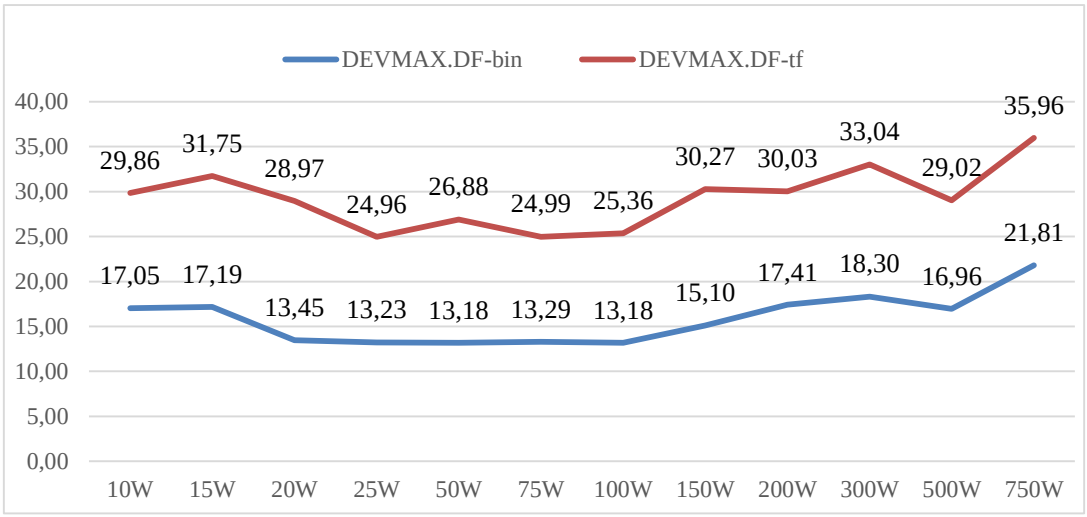
Κατά τη συσταδοποίηση φάνηκε ότι τα ποσοστά ICI ήταν μικρότερα με τη χρήση της DEVMAX.DF σε σχέση με την TF.IDF (Διάγραμμα 4.2-4). Ιδιαίτερα στα χαμηλά διανύσματα η διαφορά μεταξύ τους φτάνει πάνω από 10%, όπου στα 30W η DEVMAX.DF αποδίδει μέσο ICI=19,09%, ενώ η TF.IDF ICI=30,07%.



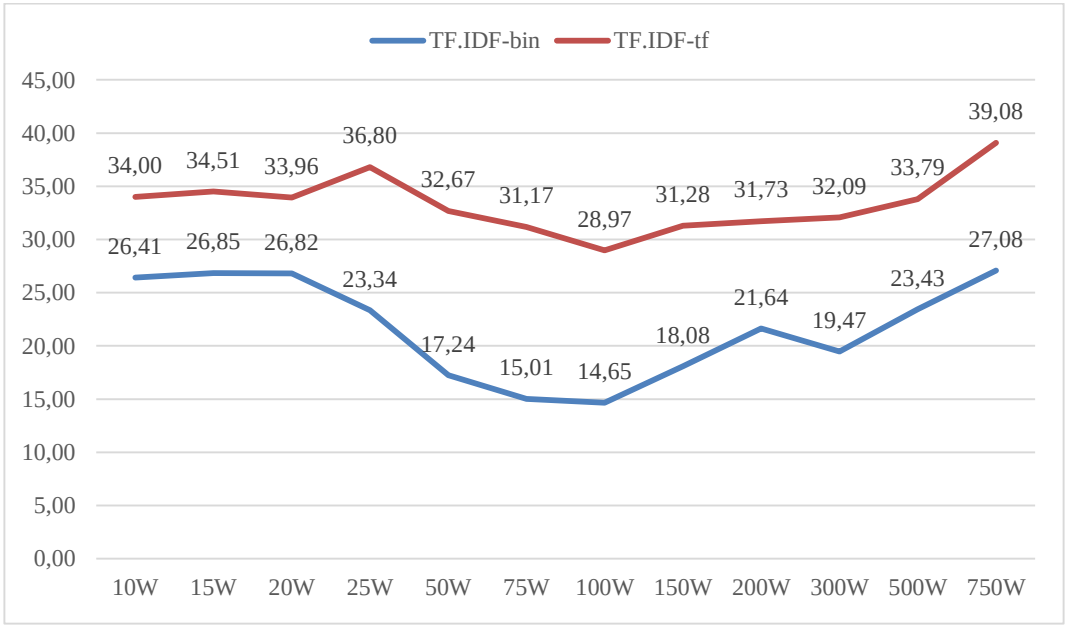
Διάγραμμα 4.2-4. Μέση τιμή ICI (%) όλων των συσταδοποιητών.

Επιπρόσθετα, η δυαδική εμφάνιση υπερίσχυσε της συχνότητας εμφάνισης. Στη χρήση της DEVMAX.DF, η δυαδική εμφάνιση λειτούργησε με σταθερό ρυθμό, ίσως

και λόγω της ήδη καλής απόδοσης της μετρικής, ρίχνοντας τα ποσοστά ICI ακόμα και στο μισό, σε αντιπαραβολή με την συχνότητα εμφάνισης (Διάγραμμα 4.2-5). Παράλληλα, πρέπει να σημειωθεί ότι η μεταχείρισή της με την TF.IDF, ενίσχυσε τη μετρική έως και 16% (75W) τα αποτελέσματα (Διάγραμμα 4.2-6).



Διάγραμμα 4.2-5. Μέση τιμή ICI(%) για τη δυαδική εμφάνιση (bin) και τη συχνότητα εμφάνισης (tf) στην DEVMAX.DF.



Διάγραμμα 4.2-6. Μέση τιμή ICI (%) για τη δυαδική εμφάνιση (bin) και τη συχνότητα εμφάνισης (tf) στην TF.IDF.

5. Συμπεράσματα

Εκείνο που προκύπτει από την έρευνα, οι τεχνικές κατηγοριοποίησης αποδίδουν ικανοποιητικά αποτελέσματα. Στην παρούσα έρευνα η κατηγοριοποίηση έφτασε στο F1=97,6% με τη χρήση του Random Forest, ενώ η συσταδοποίηση στο ICI=4,46% με τη χρήση του SOM. Κοντά σε αυτά τα ποσοστά κυμάνθηκαν και άλλοι κατηγοριοποιητές, όπως ο NBMultinomial και ο SMO, και συσταδοποιητές, όπως η EM και ο CascadeSKM. Είναι, βεβαίως, φανερό ότι και οι δύο διαδικασίες πήγαν πολύ καλά.

Αξιοσημείωτη, βεβαίως, ήταν η χρήση της μετρικής βαρύτητας DEVMAX.DF, η οποία ξεπέρασε σημαντικά την ήδη κατακεκριμένη από πολλούς TF.IDF. Επίσης, ο συνδυασμός της πρώτης με τη δυαδική εμφάνιση των όρων (DEVMAX.DF-bin) ήταν ευεργετικός. Θεωρείται ότι όλα τα αποτελέσματα συνεισφέρουν στην ανάπτυξη τεχνολογιών στο χώρο των ψηφιακών βιβλιοθηκών, ενώ επεκτείνουν τη γνώση στον τομέα της Μηχανικής Μάθησης και συμβάλλουν στις μετρικές βαρύτητας.

Η τεχνικές κατηγοριοποίησης μπορούν να αποτελέσουν σημαντικό εργαλείο για τις ψηφιακές βιβλιοθήκες, «λύνοντας τα χέρια» στους επιστήμονες της πληροφόρησης. Η πρακτική εφαρμογή πάνω στην ταξινόμηση δεν στέκεται μόνο στο κείμενο. Η Μηχανική Μάθηση πλέον παρέχει τεχνικές για την κατηγοριοποίηση και άλλων μορφών τεκμηρίων, όπως ήχο, εικόνα και βίντεο.

Ακόμα, θα μπορούσε να υποδείξει νέες δομές και συνδέσεις στο χάρτη της γνώσης. Ειδικά στην εποχή μας όπου η διεπιστημονικότητα και η συνεργία μεταξύ των τομέων οδηγεί σε μεγάλη ανάπτυξη, η χρήση της μπορεί να ωθήσει στην ανακάλυψη νέων συνεργατικών πεδίων στην επιστήμη.

Επιπλέον, η βιβλιογραφία βρίθει από έρευνες που χρησιμοποιούν τους αλγόριθμους Μηχανικής Μάθησης για οποιαδήποτε είδους σχετική εργασία σε διάφορα είδη δεδομένων, όπως είναι τα ιατρικά δεδομένα. Παραδείγματος χάριν, ένα ιατρικό αποθετήριο θα μπορούσε να τους εκμεταλλευτεί, ιδιαίτερα αν αναλογιστεί κανείς ότι οι ίδιες τεχνικές χρησιμοποιούνται και στην εξόρυξη δεδομένων.

Αβίαστα, λοιπόν, συνεπάγεται η σημασία της εισαγωγής των εργασιών αυτόματης κατηγοριοποίησης στο χώρο των βιβλιοθηκών για τη βελτίωσή τους. Οι μελλοντικές έρευνες πάνω στο θέμα, θα μπορούσαν να ασχοληθούν με την κατηγοριοποίηση άλλων μορφών ψηφιακών τεκμηρίων, ή ακόμα και τον αντίκτυπο που θα μπορούσαν να είχαν στην παραγωγικότητα μιας βιβλιοθήκης.

6. Βιβλιογραφία

Ξενόγλωσση

Abraham, A. (2005). "Artificial neural networks". *Handbook of measuring system design*, pp. 901-908. [τεκμήριο pdf, URL: http://wsc10.softcomputing.net/ann_chapter.pdf, ημερομηνία πρόσβασης: 26.7.2016].

Alpaydin, E. (2010). *Introduction to machine learning.* The MIT Press.

Arora, R., & Suman (2012). "Comparative analysis of classification algorithms on different datasets using WEKA". *International Journal of Computer Applications*, 54(13). [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9202&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 26.7.2016].

arXiv. (2016). *"arXiv.org e-Print archive".* [τεκμήριο html, URL: https://arxiv.org/, ημερομηνία πρόσβασης: 26.6.2016].

Awad, W. A., & ELseuofi, S. M. (2010). "Machine learning methods for spam e-mail classification". *International Journal of Computer Science & Information Technology (IJCSIT),* 3(1), pp. 173-184. [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.3185&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 26.7.2016].

Breinman, L., et al. (1984). *Classification and regression trees*. Pacific Grove: Wadsworth & Brooks.

Bresfelean, V. P. (2007). "Analysis and predictions on students' behavior using decision trees in Weka environment". In: *Proceedings of the ITI*, pp. 25-28. [τεκμήριο pdf, URL: http://hnk.ffzg.hr/bibl/iti2007/101%20Special%20Session/101-04-060.pdf, ημερομηνία πρόσβασης: 28.7.2016].

Bin Othman, M. F., & Yau, T. M. S. (2007). "Comparison of different classification techniques using WEKA for breast cancer". In: *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006*, pp. 520-523. [τεκμήριο pdf, URL: http://www.cs.odu.edu/~mukka/cs795sum10dm/Lecturenotes/Day3/fulltext.pdf, ημερομηνία πρόσβασης: 10.7.2016].

Bouckaert, R. R., et al. (2010). "WEKA- experiences with a Java open-source project". *Journal of Machine Learning Research*, 11, pp. 2533-2541 [τεκμήριο pdf, URL: http://www.cs.waikato.ac.nz/~ml/publications/2010/bouckaert10a.pdf, ημερομηνία πρόσβασης: 17.7.16].

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences.* MIT press.

Candela, L., Castelli, D., & Pagano, P. (2011). "History, evolution, and impact of digital libraries". In: Iglezakis, I., Synodinou, T. E., Kapidakis, S. (eds), *E-publishing and digital libraries: legal and organizational issues*. pp. 1-30. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Leonardo_Candela/publication/229422428_History_Evolution_and_Impact_of_Digital_Libraries/links/09e415110d91b1c40f000000.pdf, ημερομηνία πρόσβασης: 26.6.2016].

Cavnar, W. B., & Trenkle, J. M. (1994). "N-gram-based text categorization". *Ann Arbor MI*, 48113(2), pp. 161-175. [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.3248&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 26.6.2016].

Charalampopoulos, I., & Anagnostopoulos, I. (2011). "A comparable study employing weka clustering/classification algorithms for web page classification". In: *(PCI), 2011 15th Panhellenic Conference on Informatics.* [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Ioannis_Anagnostopoulos/publication/233991845_WEKA_Web_page_classification_draft/links/0fcfd50ddb26734710000000.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Chelba, C., Hazen, T. J., & Saraclar, M. (2008). "Retrieval and browsing of spoken content". *IEEE Signal Processing Magazine*, 25(3), pp. 39-49. [τεκμήριο pdf, URL: http://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/08-05_Hazen_IEEE-JA-13065.pdf, ημερομηνία πρόσβασης: 15.6.2016].

Cleary, J. G., & Trigg, L. E. (1995). "K*: An instance-based learner using an entropic distance measure". In: *Proceedings of the 12th International Conference on Machine learning*, pp. 108-114. [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=372B98E9DBDF6F6AF4C4F1D51F9585F8?doi=10.1.1.51.4098&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 15.6.2016].

Cohen, H., & Lefebvre, C. (Eds.). (2005). *Handbook of categorization in cognitive science.* Elsevier.

Copeland, M. (2016). "The Difference Between AI, Machine Learning, and Deep Learning?". *NVIDIA Blog*. [τεκμήριο html, URL: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/, ημερομηνία πρόσβασης: 30.6.2016].

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: information retrieval in practice.* Addison-Wesley.

Do, C. B., & Batzoglou, S. (2008). "What is the expectation maximization algorithm?". *Nature* biotechnology, 26(8), pp. 897-899. [τεκμήριο html, URL:

http://www.nature.com/nbt/journal/v26/n8/full/nbt1406.html, ημερομηνία πρόσβασης: 30.10.2016].

Efron, M. (2010). "Hashtag retrieval in a microblogging environment". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. [τεκμήριο pdf, URL: http://people.ischool.illinois.edu/~mefron/papers/efron-sigir2010.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Fensel, D. (2001). "Ontologies". *Ontologies*, pp. 11-18. [τεκμήριο pdf, URL: http://software.ucv.ro/~cbadica/didactic/sbc/documente/silverbullet.pdf, ημερομηνία πρόσβασης: 20.7.2016].

Frank, E., et al. (2004). "Data mining in bioinformatics using Weka". *Bioinformatics*, 20(15), pp. 2479-2481. [τεκμήριο pdf, URL: http://bioinformatics.oxfordjournals.org/content/20/15/2479.full.pdf, ημερομηνία πρόσβασης: 26.10.2016].

Fürnkranz, J. (1999). "Separate-and-conquer rule learning". *Artificial Intelligence Review*, 13(1), pp. 3-54. [τεκμήριο pdf, URL: http://yaroslavvb.com/papers/furnkranz-separate.pdf, ημερομηνία πρόσβασης: 16.7.2016].

Ginsparg, P. (1994). "First steps towards electronic research communication". *Computers in physics*, 8(4), pp. 390-396. [τεκμήριο pdf, URL: http://scitation.aip.org/content/aip/journal/cip/8/4/10.1063/1.4823313 πρόσβασης: 20.10.2016].

Hall, M. et al. (2009). "The WEKA data mining software: An update". *ACM SIGKDD explorations newsletter*, 11(1), pp. 10-18. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Mark_Hall6/publication/221900777_The_WEKA_data_mining_software_An_update/links/09e41507f01ad2a029000000.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Hjørland, B. (2012). "Is classification necessary after Google?", *Journal of Documentation*, 68(3), pp. 299-317 [τεκμήριο html, URL: http://dx.doi.org/10.1108/00220411211225557, ημερομηνία πρόσβασης: 19.6.2016]

IFLA, & UNESCO. (2014). "Manifesto for Digital Libraries". In: *36th session of the General Conference of UNESCO, October 25- November 10* [τεκμήριο pdf, URL: http://www.ifla.org/files/assets/digital-libraries/documents/ifla-unesco-digital-libraries-manifesto.pdf, ημερομηνία πρόσβασης: 20.6.2016].

Irani, D., et al. (2010). "Study of trend-stuffing on twitter through text classification". In: *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*. [τεκμήριο pdf, URL:

https://www.researchgate.net/profile/Calton_Pu/publication/228632556_Study_of_Trend-Stuffing_on_Twitter_through_Text_Classification/links/0c96052156388217530000000.pdf, ημερομηνία πρόσβασης: 20.6.2016].

John, G. H., & Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers". In: *11th Conference on Uncertainty in Artificial Intelligence, San Mateo*, pp. 338-345. [τεκμήριο pdf, URL: http://web.cs.iastate.edu/~honavar/bayes-continuous.pdf, ημερομηνία πρόσβασης: 18.7.2016].

Jones, K. S. (1972). "A statistical interpretation of term frequency and its application in retrieval". *Journal of Documentation*, 28(1), pp. 11-21. [τεκμήριο pdf, URL: https://pdfs.semanticscholar.org/4f09/e6ec1b7d4390d23881852fd7240994abeb58.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Joshi, A. P. & Panchal, R. K. (2014). "Data mining for staff recruitment in education system using WEKA". *International journal of research in computer science and management*, 2(1), pp. 1-4. [τεκμήριο pdf, URL: http://jspm.edu.in/InternationalJournal/joshi.pdf, ημερομηνία πρόσβασης: 20.7.2016].

Kim, S. B., et al. (2006). "Some effective techniques for naive bayes text classification". *IEEE transactions on knowledge and data engineering*, 18(11), pp 1457-1466. [τεκμήριο pdf, URL: http://ir.kaist.ac.kr/papers/2006/some%20effective%20techniques%20for%20naive%20bayes%20text%20classification.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *International Joint Conference on Artificial Intelligence (IJCAI)*, pp.1137-1145. [τεκμήριο pdf, URL: https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf, ημερομηνία πρόσβασης: 10.6.2016].

Lan, M., Tan, C. L., & Low, H. B. (2006). "Proposing a new term weighting scheme for text categorization". *AAAI*, 6, pp. 763-768. [τεκμήριο pdf, URL: http://www.aaai.org/Papers/AAAI/2006/AAAI06-121.pdf, ημερομηνία πρόσβασης: 5.6.2016].

Lan, M., et al. (2009). "Supervised and traditional term weighting methods for automatic text categorization". *IEEE transactions on pattern analysis and machine intelligence*, 31(4), pp. 721-735. [τεκμήριο pdf, URL: https://www-old.comp.nus.edu.sg/~tancl/publications/j2009/PAMI2007-v3.pdf, ημερομηνία πρόσβασης: 4.6.2016].

Machine Learning Group at the University of Waikato. (2016). *"Weka 3: Data Mining Software in Java"*. [τεκμήριο html, URL: http://www.cs.waikato.ac.nz/ml/weka/index.html, ημερομηνία πρόσβασης: 17.7.2016].

Marr, B. (2016). "A Short History of Machine Learning -- Every Manager Should Read:. *Forbes*. [τεκμήριο html, URL: http://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#13627e6a323f, ημερομηνία πρόσβασης: 30.10.2016].

McCallum, A., & Nigam, K. (1998). "A Comparison of Event Models for Naive Bayes Text Classification". In: *AAAI-98 Workshop on 'Learning for Text Categorization*. [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 17.7.2016].

McCallum, A., Nigam, K., & Ungar, L. H. (2000). "Efficient clustering of high-dimensional data sets with application to reference matching". In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 169-178. [τεκμήριο pdf, URL: http://www.kamalnigam.com/papers/canopy-kdd00.pdf, ημερομηνία πρόσβασης: 17.7.2016].

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

Nilsson, N. J. (1996). *Introduction to machine learning. An early draft of a proposed textbook.*

Peel, D., & McLachlan, G. J. (2000). "Robust mixture modelling using the t distribution". *Statistics and computing*, 10(4), pp. 339-348. [τεκμήριο pdf, URL: https://pdfs.semanticscholar.org/99ad/63b1c7a70e2022ed0f82671ed1e3e8ef2772.pdf, ημερομηνία πρόσβασης: 25.7.2016].

Peterson, L. E. (2009). "K-nearest neighbor". *Scholarpedia*, 4(2). [τεκμήριο html, URL: http://scholarpedia.org/article/K-nearest_neighbor, ημερομηνία πρόσβασης: 25.7.2016].

Pomerantz, J., & Marchionini, G. (2007) "The digital library as place". *Journal of Documentation*, (63)4, pp. 505-533. [τεκμήριο html, URL: http://dx.doi.org/10.1108/00220410710758995, ημερομηνία πρόσβασης: 20.7.2016].

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). "Cultural route to the emergence of linguistic categories". *PNAS*, 105(23), pp. 7936-7940. [τεκμήριο html, URL: http://www.pnas.org/content/105/23/7936.long, ημερομηνία πρόσβασης: 25.7.2016].

Quinlan, J. R. (1986). "Induction of decision trees". *Machine learning*, 1(1), pp. 81-106. [τεκμήριο pdf, URL: http://grc.mnnu.edu.cn/~fmin/materials/course/roughsets%5Creferences%5CQuinlanJR1986_InductionOfDecisionTrees.pdf, ημερομηνία πρόσβασης: 23.6.2016].

Ramos, J. (2003). "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning.* [τεκμήριο pdf, URL: https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf, ημερομηνία πρόσβασης: 14.6.2016].

Rajput, A., et al. (2011). "J48 and JRIP rules for e-governance data". *International Journal of Computer Science and Security (IJCSS)*, 5(2), pp. 201. [τεκμήριο pdf, URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.4477&rep=rep1&type=pdf, ημερομηνία πρόσβασης: 25.7.2016].

Rish, I. (2001). "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3(22) pp. 41-46. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Irina_Rish/publication/228845263_An_Empirical_Study_of_the_naive_Bayes_Classifier/links/00b7d52dc3ccd8d692000000.pdf, ημερομηνία πρόσβασης: 25.7.2016].

Ruef, M., & Patterson, K. (2009). "Credit and classification: The impact of industry boundaries in nineteenth-century America". *Administrative Science Quarterly*, 54(3), pp. 486-520. [τεκμήριο pdf, URL: http://www.edegan.com/pdfs/Ruef%20Patterson%20(2009)%20-%20Credit%20And%20Classification%20The%20Impact%20Of%20Industry%20Boundaries%20On%20Nineteenth%20Century%20America.pdf, ημερομηνία πρόσβασης: 23.7.2016].

Sammut, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Samuel, A. L. (1959). "Some studies in machine learning using the game checkers". *IBM Journal of research and development*, 3(3), pp. 210-229. [τεκμήριο pdf, URL: https://pdfs.semanticscholar.org/e9e6/bb5f2a04ae30d8ecc9287f8b702eedd7b772.pdf, ημερομηνία πρόσβασης: 25.7.2016].

Schifanella, R., et al. (2010). "Folks in folksonomies: social link prediction from shared metadata". In: *Proceedings of the third ACM international conference on Web search and data mining*, pp. 271-280. [τεκμήριο pdf, URL: http://arxiv.org/pdf/1003.2281.pdf, ημερομηνία πρόσβασης: 24.7.2016].

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

Sembok, T. M. T., & Bakar, Z. A. (2011). "Effectiveness of stemming and n-grams string similarity matching on Malay documents". *International Journal of Applied Mathematics and Informatics*, 5(3), pp. 208-215. [τεκμήριο pdf, URL: http://www.naun.org/main/UPress/ami/20-656.pdf, ημερομηνία πρόσβασης: 20.7.2016].

Shavlik, J. W., & Dietterich, T. G. (eds.) (1990). *Readings in machine learning*. Morgan Kaufmann Publ.

Soucy, P., & Mineau, G. W. (2005). "Beyond TFIDF weighting for text categorization in the vector space model". *IJCAI*, 5, pp. 1130-1135. [τεκμήριο pdf, URL: http://s3.amazonaws.com/academia.edu.documents/38465965/0304.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1474218863&Signature=vE4qBRuv5fXGWKc6T3HEAfGDfBk%3D&response-content-disposition=inline%3B%20filename%3DBeyond_TFIDF_Weighting_for_Text_Categori.pdf, ημερομηνία πρόσβασης: 20.7.2016].

Specia, L., & Motta, E. (2007). "Integrating folksonomies with the semantic web". In: *European Semantic Web Conference*, pp. 624-639. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Lucia_Specia/publication/225466445_Integrating_Folksonomies_with_the_Semantic_Web/links/56179c2b08ae0224ebce9957.pdf, ημερομηνία πρόσβασης: 20.8.2016].

Staab, S., & Studer, R. (Eds.). (2013). *Handbook on ontologies. Springer Science & Business Media*.

Triantafyllou, I., Demiros, I., & Piperidis, S. (2001). "Two Level Self-Organizing Approach to Text Classification". *RANLP-2001: Recent Advances in NLP*.

Triantafyllou, I., et al. (2014). "Significance of Clustering and Classification Applications in Digital and Physical Libraries". In: *4th International Conference IC-ININFO, 5-8 September*.

Turcinek, P., Stastny, J., & Motycka, A. (2012). "Usage of data mining techniques on marketing research data". In: *Proceedings of the 11th WSEAS international conference on Applied Computer and Applied Computational Science*, pp. 159-164. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Jiri_Stastny/publication/262165434_Usage_of_data_mining_techniques_on_marketing_research_data/links/54b906410cf269d8cbf72a37.pdf, ημερομηνία πρόσβασης: 20.8.2016].

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 49, pp. 433-460. [τεκμήριο pdf, URL: https://www.csee.umbc.edu/courses/471/papers/turing.pdf, ημερομηνία πρόσβασης: 10.6.2016].

Vorgia, F., Triantafyllou, I., & Koulouris, A. (2017). “Hypatia Digital Library: A Text Classification Approach Based on Abstracts”. In: *Strategic Innovative Marketing*, pp. 727-733. Springer International Publishing.

Wikipedia. (2016a). *Machine learning*. [τεκμήριο html, URL: https://en.wikipedia.org/wiki/Machine_learning, ημερομηνία πρόσβασης: 25.7.2016].

Wikipedia. (2016b). *Neuron.* [τεκμήριο html, URL: https://en.wikipedia.org/wiki/Neuron, ημερομηνία πρόσβασης: 25.7.2016].

Wilbur, W. J, & Sirotkin, K. (1992). “The automatic identification of stop words”. *Journal of Information Science*, 18(1), pp. 45-55. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/W_Wilbur/publication/247786801_The_automatic_identification_of_stop_words/links/552286970cf29dcabb0d721a.pdf, ημερομηνία πρόσβασης: 15.7.2016].

Witten, I. H., Frank, E., & Hall, A. M. (2011). *Data mining: Practical Machine Learning Tools and Techniques.* Elsevier.

Yadav, A. K., Malik, H., & Chandel, S. S. (2014). “Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models”. *Renewable and Sustainable Energy Reviews*, 31, pp. 509-519. [τεκμήριο pdf, URL: https://www.researchgate.net/profile/Amit_Yadav27/publication/260015601_Selection_of_most_relevant_input_parameters_using_WEKA_for_artificial_neural_network_based_solar_radiation_prediction_models/links/00b7d53ac2c807f85b000000.pdf, ημερομηνία πρόσβασης: 15.7.2016].

Yang, Y., & Liu, X. (1999). “A re-examination of text categorization methods”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42-49. [τεκμήριο pdf, URL: http://people.csail.mit.edu/jim/temp/yang.pdf, ημερομηνία πρόσβασης: 15.7.2016].

Ελληνική

Καπιδάκης, Σ. (2014). *Εισαγωγή στις ψηφιακές βιβλιοθήκες.* Θεσσαλονίκη: Εκδόσεις Δίσιγμα.

Κυριάκη-Μάνεση, Δ., & Κουλούρης, Α. (2015). *Διαχείριση ψηφιακού περιεχομένου.* Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. [τεκμήριο ηλεκτρ. βιβλ. URL: http://hdl.handle.net/11419/2496, ημερομηνία πρόσβασης: 20.6.2016].