Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Ανάλυση Επιστημονομετρικών Δεδομένων με Μεθόδους Μηχανικής Μάθησης

Διπλωματική Εργασία

ΑΝΤΩΝΙΟΥ ΜΑΡΕΛΑ

Επιβλέπων
Μιχαήλ Βασιλακόπουλος
Αναπληρωτής Καθηγητής

Βόλος, Ιούνιος 2019

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Ανάλυση Επιστημονομετρικών Δεδομένων με Μεθόδους Μηχανικής Μάθησης

Διπλωματική Εργασία

ΑΝΤΩΝΙΟΥ ΜΑΡΕΛΑ

Επιτροπή επίβλεψης

Επιβλέπων	Συνεπιβλέπουσα	Συνεπιβλέπουσα
Μιχαήλ Βασιλακόπουλος	Ελένη Τουσίδου	Παναγιώτα Τσομπανοπούλου
Αναπληρωτής Καθηγητής	Μέλος Ε.ΔΙ.Π.	Αναπληρώτρια Καθηγήτρια

Βόλος, Ιούνιος 2019

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Ο/Η συγγραφέας αυτής της εργασίας βεβαιώνει ότι κάθε βοήθεια την οποία είχε για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης βεβαιώνει ότι έχει αναφέρει τις όποιες πηγές από τις οποίες έκανε χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται επακριβώς, είτε παραφρασμένες.

University of Thessaly
Faculty of Engineering
Department of Electrical & Computer Engineering

Scientometric Data Analysis using Machine Learning Methods

Diploma Thesis

ANTONIOS MARELAS

Supervisor
Michael Vassilakopoulos
Associate Professor

Volos, June 2019

Περίληψη

Τα τελευταία χρόνια, ο όγκος των δημοσιευμένων ερευνών αλλά και το ποσοστό των ερευνητών, έχει μεγαλώσει ραγδαία. Η επιστημονική πολυμορφία που διέπει τον κλάδο της επιστημονομετρίας, ολοένα και αυξάνεται. Το περιβάλλον πλέον, είναι περισσότερο ανταγωνιστικό και οι συνθήκες ανάδειξης αξιόπιστων επιστημονικών ερευνών διέπονται από πολυδιάστατους παράγοντες. Υπάρχει η ανάγκη για βελτίωση της ανάλυσης και μελέτης της δυναμικής της επιστήμης, με απώτερο σκοπό να αντιμετωπιστούν οι καινούριες προκλήσεις και οι τάσεις της εποχής. Ο τομέας της επιστημονομετρίας και άλλων συναφών επιστημονικών πεδίων, όπως η βιβλιομετρία, διαδραματίζουν σημαντικό ρόλο στη διαδικασία εκτίμησης του επιστημονικού δυναμικού. Οι δυνατότητες και τα εργαλεία που διαθέτουν οι προαναφερόμενοι κλάδοι, συντέλεσαν στη ραγδαία ανάπτυξη τους. Το ανταγωνιστικό περιβάλλον στις μέρες μας, καθιστά την ανάγκη για εκπόνηση και συγγραφή μίας επιστημονικής εργασίας που θα χαρακτηρίζεται από σαφήνεια, καινοτομία και ποιοτική τεκμηρίωση των παρουσιασθέντων αποτελεσμάτων της. Ανέκαθεν οι συγγραφείς αναζητούσαν τρόπους για την αναγνώριση των ερευνών τους και την απήχησή τους στον κόσμο της επιστήμης. Οι παράγοντες που καθορίζουν τη μελλοντική αναγνώριση των ερευνών είναι αρκετοί και σίγουρα δεν μπορούν να ιεραρχηθούν με βάση την επιρροή τους. Το αντίκτυπο μίας έρευνας, συνοδεύεται από ένα συνοθύλευμα κριτηρίων που πρέπει να ληφθούν υπ' όψιν, προκειμένου να λάβει μελλοντικές αναφορές και να αναγνωρισθεί τόσο από την επιστημονική όσο και από την ευρεία κοινότητα. Έτσι, λόγω της ανάγκης για ποσοτικοποίηση του αντίκτυπου των ερευνών, επικεντρωθήκαμε στην εξαγωγή κάποιων πειραμάτων με τη χρήση μεθόδων της Μηχανικής Μάθησης (Μ.Μ), για τον εντοπισμό πληροφοριών από ένα σύνολο δημοσιευμένων επιστημονικών εγγράφων. Αρχικά, παρουσίαζεται η διαδικασία μοντελοποίησης θεμάτων που πραγματεύονται τα έγγραφα, χρησιμοποιώντας έναν αλγόριθμο που πετυχαίνει ικανοποιητικά αποτελέσματα, μετά από διαφορετικά πειράματα που αφορούν τις παραμέτρους του. Εν συνεχεία, πραγματοποιείται η διαδικασία της ομαδοποίησης εγγράφων, δηλαδή η ταξινόμηση εγγράφων με ίδια χαρακτηριστικά, στην ίδια ομάδα, αναπτύσσοντας μέσω ενός αλγορίθμου, 2 μοντέλα. Το ένα μοντέλο ομαδοποιεί έγγραφα που "μοιράζονται" παρόμοια θέματα που προέκυψαν από τη διαδικασία της μοντελοποίησης και το άλλο εστιάζει στην ομαδοποίηση εγγράφων, με χαρακτηριστικά τις αναφορές που έλαβαν εντός συγκεκριμένης χρονικής περιόδου και το μέγεθος των τίτλων τους. Στο τελευταίο μοντέλο συσταδοποίησης, χρησιμοποιήθηκε μία τεχνική που βρίσκει το βέλτιστο αριθμό συστάδων. Δεδομένου των διαφορετικών χαρακτηριστικών που διέπουν τα διάφορα σύνολα επιστημονομετρικών δεδομένων, αυτό το μοντέλο αναδεικνύει την ανάγκη για συνδυασμό των παραγόντων που συντελούν στη μελλοντική επιτυχία. Τέλος, αναπτύσσεται μετά από μελέτη και παρουσίαση πει-

ραμάτων αξιολόγησης, ένα νευρωνικό δίκτυο που προβλέπει τις μελλοντικές αναφορές που θα λάβει ένα έγγραφο την επόμενη τριετία. Το νευρωνικό πετυχαίνει σύγκλιση αρκετά γρήγορα και η ακρίβεια πρόβλεψης του μεταξύ των προβλεπόμενων και των πραγματικών τιμών στα δεδομένα εκπαίδευσης, αγγίζει έως και το 99%.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Επιστημονομετρία, Μοντελοποίηση θεμάτων, Συσταδοποίηση, Πρόβλεψη Μελλοντικών αναφορών, Νευρωνικά Δίκτυα, Python

Abstract

In recent years, the volume of published research as well as the percentage of researchers have grown rapidly. Scientific diversity, which governs the science sector, is increasing. The environment is now more competitive and the conditions for credible scientific research are governed by multidimensional factors. There is a need to improve the analysis and study of the dynamics of science, with a view to addressing the new challenges and trends of the time. The field of scientometrics and other relevant scientific fields such as bibliometrics, play an important role in the process of assessing scientific potential. The capabilities and tools of the aforementioned sectors have contributed to their rapid growth. The competitive environment nowadays, makes the need for writing and writing a scientific work, characterized by clarity, innovation and quality documentation of the results presented. The writers have always been seeking ways to recognize their research and their impact on the world of science. The factors that determine the future recognition of investigations are numerous and certainly can not be prioritized based on their influence. The impact of an investigation is accompanied by a mix of criteria to be taken into account in order to receive future citations and be recognized by both the scientific community and the broad community. Thus, due to the need to quantify the impact of surveys, we focused on the extraction of some experiments using Machine Learning methods to identify information from a set of published scientific papers. Initially, the process of modeling issues dealing with documents is presented using an algorithm that achieves satisfying results after different experiments involving its parameters. Then, the document clustering process is performed which is the classification of documents with the same attributes, in the same group, by developing 2 models by means of an algorithm. One model groups documents that share the same themes that arose from the modeling process, while the other focuses grouping of documents with features of the citations they received within a specific time period and the size of their titles. In the latest clustering model, a technique has been used that finds the optimal number of clusters. Given the different characteristics of the different sets of scientific data, this model highlights the need for a combination of factors contributing to future success. Finally, it develops after studying and presenting assessment experiments, a neural network that predicts future citations that a document will receive in the next three years. The neural achieves convergence quite quickly and its prediction accuracy between the predicted and actual values in the training data, reaches up to 99%.

Keywords

Machine Learning, Scientometrics, Topic Modeling, Clustering, Prediction of Future Citations, Neural Networks, Python

Στην οικογένειά μου.

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Βασιλακόπουλο Μιχαήλ, για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω, παρέχοντας μου πολύτιμη βοήθεια και καθοδήγηση καθ' όλη τη διάρκεια της συγγραφής της. Επίσης, θα ήθελα να ευχαριστήσω θερμά τις δύο συνεπιβλέπουσες καθηγήτριες, την κ. Τσομπανοπούλου Παναγιώτα, αναπληρώτρια καθηγήτρια του Πανεπιστημίου Θεσσαλίας και την κ. Τουσίδου Ελένη, μέλος Ε.ΔΙ.Π, για την πολύτιμη βοήθειά τους στη διαδικασία αξιολόγησης της παρούσας διπλωματικής. Επιπλέον, σημαντική βοήθεια και στήριγμα όλα αυτά τα χρόνια ήταν οι φίλοι μου, μέσα από τους οποίους κατάφερα πολλές φορές να αντιμετωπίσω δύσκολες καταστάσεις και συνθήκες. Τέλος, θα ήθελα να πω ένα μεγάλο ευχαριστώ στους σημαντικότερους ανθρώπους της ζωής μου, τους γονείς μου και τα αδέρφια μου, για την ηθική συμπαράσταση και τον αγώνα τους όλα αυτά τα χρόνια.

Πρόλογος

Η παρούσα διπλωματική εργασία, εκπονήθηκε ως το τελευταίο βήμα για την απόκτηση του διπλώματος του τμήματος Ηλεκτρολόγων Μηχανικών & Μηχανικών Ηλεκτρονικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας, στην πόλη του Βόλου. Η εργασία πραγματοποιήθηκε υπό την επίβλεψη του αναπληρωτή καθηγητή του Πανεπιστημίου Θεσσαλίας, του κ. Βασιλακόπουλου Μιχαήλ. Η αρχική ενασχόληση με τον τομέα που εστιάζει η παρούσα εργασία, δηλαδή με το κομμάτι της Τεχνητής Νοημοσύνης και κυρίως της Βαθιάς Μάθησης, ήταν με την επιλογή κάποιων μαθημάτων στο Πανεπιστήμιο Θεσσαλίας, που σχετίζονται με τον προαναφερόμενο τομέα. Η ολοένα και περισσότερο αυξανόμενη ανάγκη για τον αυτόματο προγραμματισμό των μηχανών, με ώθησαν ακόμη περισσότερο προς τον κλάδο της Τεχνητής Νοημοσύνης και μου προκάλεσαν το ενδιαφέρον. Γενικότερα, ο προγραμματισμός για την αυτοματοποίηση και τη διευκόλυνση του επιπέδου ζωής, σε συνδυασμό με τη δυνατότητα για εξαγωγή πληροφοριών από αδόμητης μορφής δεδομένα, ήταν κάτι που από την πρώτη στιγμή μου έδωσε το κίνητρο για την ενασχόληση μου με αυτό το κομμάτι. Συγκεκριμένα, το ερέθισμα για το θέμα της εκπονηθείσας διπλωματικής, το έλαβα από την ανάγνωση κάποιων άρθρων που σχετίζονταν με την ανάλυση της δυναμικής της επιστήμης.

Την τελευταία δεκαετία, η ύπαρξη αποθετηρίων μεγάλων επιστημονομετρικών δεδομένων (big scholarly data), προσφέρουν τη δυνατότητα για εφαρμογή πολλών μεθόδων ανάλυσης στον τομέα της επιστημονομετρίας, με απώτερο σκοπό τη μελέτη και κατασκευή μοντέλων, εξάγοντας πληροφορίες που προηγουμένως δεν ήταν εμφανής. Έτσι, θεώρησα ότι ήταν μία καλή ευκαιρία, έχοντας θεωρητικό και πρακτικό υπόβαθρο από τα μαθήματα της σχολής, να αναπτύξω κάποια πειράματα με την κατασκευή ανάλογων μοντέλων, ώστε να μπορέσω να εντοπίσω και εγώ με την σειρά μου, την προσφορά της Μηχανικής Μάθησης στον κλάδο της επιστημονομετρίας. Ο κλάδος αυτός, αποτελεί την κινητήρια δύναμη για την ανάλυση και αξιολόγηση δημοσιευμένων ερευνών από διάφορους επιστημόνες και αποτέλεσε τη βάση της παρούσας εργασίας.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Περιγραφή του Προβλήματος

Τα τελευταία χρόνια, ο κλάδος της επιστημονομετρίας είναι από τους πιο πολύτιμους κλάδους για την ανάλυση της επιστήμης και της ερευνητικής διαδικασίας. Το σύνολο της δημοσιευμένης βιβλιογραφίας έχει αυξηθεί ραγδαία και ολοένα και περισσότεροι συγγραφείς δημοσιεύουν εργασίες και έρευνες, δημιουργώντας έτσι δύσκολες συνθήκες για την ανάλυση όλων αυτών των επιστημονομετρικών δεδομένων. Σήμερα, πολλοί είναι οι ερευνητές που εργάζονται σε επιστημονικά έργα και δημοσιεύουν έρευνες καθημερινά. Η επιτακτική ανάγκη για αποτελεσματική αξιολόγηση όλων των ερευνητικών εργασιών καθίσταται απαραίτητη. Η διάκριση ερευνών που διέπονται από σημαντικά ή και όχι τόσο σημαντικά αποτελέσματα καθώς και η απήχηση αυτών στον επιστημονικό τομέα, έχει παροτρύνει τον επιστημονικό κόσμο και τους αναλυτές, να ποσοτικοποιούν με διάφορα μέτρα, την επιρροή που έχει ένα δημοσιευμένο έγγραφο.

Η διαδικασία εξαγωγής σημαντικών πληροφοριών καθορίζεται από πολλούς παράγοντες. Ο κυριότερος παράγοντας που χρησιμοποιείται ευρέως για να χαρακτηριστεί μία έρευνα αξιόπιστη, είναι ο αριθμός των παραπομπών που έχει λάβει. Αυτός ο παράγοντας αποτελεί και τη βάση για πολλούς άλλους παράγοντες, όπως ο h-index και άλλους που θα αναλύσουμε στο Κεφάλαιο 3. Βέβαια, η ποικιλομορφία των δημοσιευμένων ερευνών σε ιδιότητες και επιπτώσεις αλλά και το ολοένα και αυξανόμενο ανταγωνιστικό περιβάλλον που παρατηρείται, δυσκολεύει τους ερευνητές να ιεραρχήσουν με απόλυτη σιγουριά τους παράγοντες αξιολόγησης των ερευνών. Το πλήθος των συνολικών αναφορών μίας έρευνας, σε συνδυασμό με άλλους παράγοντες που διαθέτουν εξίσου σημαντικό ρόλο στην απήχηση αυτής, αποτελούν οδηγό για τον εντοπισμό του έργου αλλά και του επιστημονικού της αντίκτυπου που πραγματοποιήθηκε στα πλαίσια της επιστημονικής κοινότητας. Διαθέτοντας λοιπόν τη δύναμη που προσφέρει η Μ.Μ στον τομέα της επιστημονομετρίας, μπορούμε και εξαγάγουμε συμπεράσματα με λιγότερο επίπονη διαδικασία, για έναν μεγάλο όγκο δημοσιευμένης βιβλιογραφίας. Πλέον, τα επιστημονομετρικά δεδομένα συνοψίζονται σε πληροφορίες που αποτελούν τα θεμέλια για την εύρεση του αντίκτυπου των επιστημονικών ερευνών, αναγκάζοντας το πρόβλημα της επιστημονομετρικής ανάλυσης να γίνεται ευκολότερο.

1.2 Στόχος της Εργασίας

Στόχος είναι η μελέτη και η ανάλυση επιστημονομετρικών δεδομένων με μεθόδους της Μ.Μ. Η εξαγωγή πληροφοριών που δεν είναι εμφανής αρχικά, είναι και ο λόγος που κάνει τη Μ.Μ να διεισδύει στον τομέα της επιστημονομετρίας. Συγκεκριμένα, θα προσπαθήσουμε να εξαγάγουμε κάποια νέα χαρακτηριστικά, από ήδη υπάρχοντα δεδομένα, αναπτύσσοντας διαφορετικά μοντέλα. Αξίζει να σημειωθεί, ότι τα αποτελέσματα των μοντέλων δίνουν τη δυνατότητα της απόκτησης πληροφοριών που είναι σημαντικές για τα εξεταζόμενα δημοσιευμένα έγγραφα και πραγματοποιείται η ανάλυση της επιστημονομετρίας και των παραγόντων της, με ένα διαφορετικό τρόπο.

Το μοντέλο πρόβλεψης μελλοντικών αναφορών χρήζει ιδιαίτερης προσοχής, καθώς το σύνολο των αναφορών αποτελεί και έναν από τους κυριότερους λόγους που μία έρευνα αποτελεί αντικείμενο μελέτης και ανάλυσης, καθώς αναμένεται να είναι και αυτή με τα πιο σημαντικά και αξιόπιστα αποτελέσματα. Η χρήση νευρωνικών δικτύων, κατέστησε ικανό το μοντέλο να προβλέπει τις αναφορές της επόμενης τριετίας, έχοντας στη διάθεσή του 12 χρονιές που αναφέρονται στο πλήθος των παραπομπών που έλαβαν τα έγγραφα. Η παρούσα εργασία, δεν στηρίζεται μόνο στις μελλοντικές αναφορές που θα λάβει ένα έγγραφο. Επιπλέον κίνητρο για την εκπόνηση περαιτέρω πειραμάτων, ήταν ο στόχος για την οργάνωση των αρχικών δεδομένων που χαρακτηρίζονται συνήθως από την απουσία κάποιας συγκεκριμένης δομής. Ο εντοπισμός λέξεων-κλειδιών αλλά και θεμάτων που διέπουν ένα σύνολο δημοσιευμένων εγγράφων, είναι εξίσου σημαντικός με τη μελλοντική πρόβλεψη αναφορών. Επιπρόσθετα, η συσταδοποίηση εγγράφων με όμοια χαρακτηριστικά, προσφέρει επίσης οργάνωση στα δεδομένα μας και εντοπίζει τις σχέσεις αυτών με τα χαρακτηριστικά τους.

Η συγκέντρωση των δεδομένων μας πραγματοποιήθηκε από περιβάλλοντα πραγματικών συνθηκών που αφορούν μία συγκεκριμένη χρονική περίοδο, καθιστώντας τα αποτελέσματα των πειραμάτων μας, πιο αξιόπιστα. Τα μοντέλα δεν αναπτύχθηκαν μέσα από πειραματικά και προσεγγιστικά δεδομένα όπως συνηθίζεται. Σε κάθε πείραμα, τα μοντέλα λαμβάνουν ως είσοδο ένα φιλτραρισμένο υποσύνολο των αρχικών δεδομένων, με απώτερο στόχο την καλύτερη παρουσίαση αποτελεσμάτων του ενίοτε μοντέλου. Έτσι, ο χρήστης είναι ικανός να κατανοήσει πλήρως όχι μόνο τη φύση των πειραμάτων, αλλά και τη φιλοσοφία πίσω από τα προβλήματα που συναντώνται στον τομέα της επιστημονομετρίας.

1.3 Οργάνωση του Τόμου

- Στο **Κεφάλαιο 1**, υπάρχει η εισαγωγή της εργασίας, όπου αναλύεται το πρόβλημα που δημιουργείται στον κλάδο της επιστημονομετρίας και δίνεται μία περιγραφή του στόχου της παρούσας διπλωματικής.

- Στο **Κεφάλαιο 2**, πραγματοποιείται μία θεωρητική εισαγωγή στο πεδίο της Τεχνητής Νοημοσύνης (Τ.Ν) και στα υποπεδία της, όπως είναι η Μ.Μ, με σκοπό την κατανόηση εννοιών που σχετίζονται με τα πεδία αυτά.

- Στο **Κεφάλαιο 3**, αναλύεται ο καθαυτό κλάδος της επιστημονομετρίας, μαζί με τα μετρικά

ερευνών, αλλά και γίνεται η παράθεση άλλων συναφών επιστημών.

- Στο **Κεφάλαιο 4**, αποσαφηνίζεται όλη η απαραίτητη θεωρία των αλγορίθμων που χρησιμοποιήθηκαν, αλλά και αναλύονται εκτενέστερα έννοιες που έχουν άμεση σύνδεση με τα αναπτυχθέντα μοντέλα.
- Στο **Κεφάλαιο 5**, παρουσιάζονται τα αποτελέσματα των πειραμάτων και η σύγκριση απόδοσης των χρησιμοποιηθέντων αλγορίθμων, τροποποιώντας διάφορες παραμέτρους τους.
- Στο **Κεφάλαιο 6**, παρατίθενται τα συμπεράσματα και οι μελλοντικές επεκτάσεις της εκπονηθείσας διπλωματικής.

Κεφάλαιο 2

Βιβλιογραφική Ανασκόπηση και Θεωρητικό Υπόβαθρο

Στο παρόν κεφάλαιο είναι σκόπιμο να αναφέρουμε τον τομέα της Μ.Μ και της Τ.Ν που αποτελεί ευρύτερο πεδίο μελέτης, αλλά και κάποιες βασικές έννοιες που βρίσκονται πίσω από τους αλγορίθμους που εφαρμόστηκαν. Επίσης θα αναφερθούμε και σε κάποιες συγγενικές εργασίες/έρευνες που έχουν πραγματοποιηθεί και έδωσαν ένα θεωρητικό υπόβαθρο και ένα επιπλέον κίνητρο για την εκπόνηση κάποιων εκ των διεξαχθέντων πειραμάτων.

2.1 Βιβλιογραφική Ανασκόπηση

Η παρούσα διπλωματική, στοχεύει στην ανακάλυψη πληροφοριών που εξάγονται με βάση ένα σύνολο επιστημονομετρικών δεδομένων. Η διαδικασία αυτή, επιτυγχάνεται με τη χρήση της Μ.Μ. Οι τεχνικές που αποτελούν την τελευταία, είναι και αυτές που με τη βοήθεια τους διεξήχθησαν τα πειράματα και συντέλεσαν στην παρουσίαση και ανάλυση των πληροφοριών. Επομένως, οι έρευνες που αποτέλεσαν πηγή γνώσεων και οδηγό για την υλοποίηση κάποιων πειραμάτων, θα σχετίζονται με τον τομέα της Μ.Μ.

Όσον αφορά τη μοντελοποίηση θεμάτων, αξιοποιήθηκε ο αλγόριθμος LDA, ο οποίος ανακαλύπτει "κρυμμένα" θέματα που εντοπίζονται σ' ένα σύνολο εγγράφων. Μεταξύ άλλων αλγορίθμων ταξινόμησης, είναι ο LSI και ο pLSI. Οι 3 αυτές μέθοδοι, στηρίζονται στη μείωση διαστάσεων του συνολικού όγκου των εγγράφων και πολλές φορές θεωρούνται και ανταγωνιστικοί. Δύο από τις έρευνες που μας βοήθησαν στο να αποφασίσουμε σε ποιον αλγόριθμο θα επικεντρωθούμε, παρατίθενται παρακάτω.

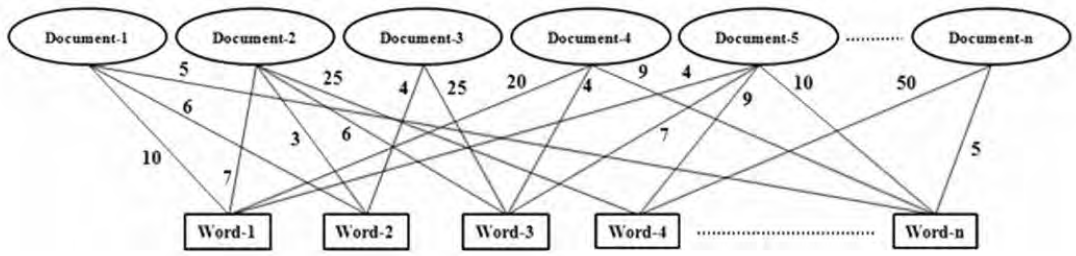
Η πρώτη έρευνα [10], στηρίχθηκε στο πρόβλημα μοντελοποίησης κειμένων και άλλων διακριτών δεδομένων και πραγματοποιήθηκε από τους David M. Blei, Andrew Y. Ng και Michael I. Jordan. Είχε ως σκοπό τον εντοπισμό περιγραφών για μέλη μίας συλλογής που να επιτρέπουν την αποτελεσματική επεξεργασία μεγάλων συλλογών, διατηρώντας ταυτόχρονα κάποιες βασικές σχέσεις στατιστικής φύσης. Στην έρευνα παρατηρείται, ότι η τεχνική pLSI μαθαίνει τα "μίγματα" των θεμάτων (κατανομή πιθανοτήτων), μόνο για τα έγγραφα στα οποία έχει προηγουμένως εκπαιδευτεί. Αυτό έχει ως αποτελέσμα, την ανικανότητα του μοντέλου να γενικεύει και να ανακαλύπτει

θέματα, διαφορετικά απ' αυτά που έχει προπονηθεί. Στην έρευνα τους αναφέρουν, ότι δεν είναι πλήρως γενετικό μοντέλο και δεν υπάρχει κανένας άμεσος τρόπος για να πραγματοποιηθεί η ανάθεση πιθανοτήτων σε έγγραφα που το μοντέλο δεν έχει ξανασυναντήσει. Ένα από τα μέτρα για την αξιολόγηση των μοντέλων ταξινόμησης είναι η σύγχυση. Έτσι, μετά από μέτρηση της σύγχυσης για διάφορα μοντέλα, συμπεριλαμβανομένου του LDA και του pLSI, απέδειξαν ότι η καμπύλη σύγχυσης του LDA, ήταν και αυτή με την αισθητά μεγαλύτερη βελτίωση, όπου τα αποτελέσματα δείχνουν ότι είναι και τα καλύτερα απ' αυτά που προκύπτουν από μοντέλα Bag-of-Words, δηλαδή μοντέλα που στηρίζονται στην εμφάνιση των πιο συχνά εμφανιζόμενων λέξεων μέσα σ' ένα έγγραφο. Επίσης, αναφέρουν ότι η τεχνική LDA, είναι ένα απλό μοντέλο και πολλές φορές, ανταγωνίζεται τις μεθόδους LSI και pLSI στη ρύθμιση της μείωσης των διαστάσεων για τις συλλογές εγγράφων. Όμως μοντέλα πιθανοτικά, όπως το μοντέλο LDA, μπορούν να κλιμακωθούν, ώστε να παρέχουν χρήσιμα μηχανήματα συμβολισμού σε τομείς που περιλαμβάνουν πολλαπλά επίπεδα δομής και χαρακτηρίζονται από επεκτασιμότητα. Ο LDA μπορεί εύκολα να ενσωματωθεί σ' ένα πιο περίπλοκο μοντέλο, ιδιότητα που δεν χαρακτηρίζει τις μεθόδους LSI και pLSI.

Η δεύτερη έρευνα [9] που είναι παρόμοια με την πρώτη, πραγματοποιήθηκε από τους Kaveh Bastani, Hamed Namavari και Jeffrey Shaffer και αφορά τα προβλήματα των καταναλωτών που αντιμετωπίζουν, βασιζόμενα σε διάφορες χρηματοπιστωτικές υπηρεσίες. Στηρίχθηκε στη μοντελοποίηση θεμάτων με τον αλγόριθμο LDA και ουσιαστικά αναζητούσαν, ένα έξυπνο σύστημα για την αυτόματη ανάλυση των παραπόνων και την παροχή μίας γενικής ιδέας στους ειδικούς. Συνέκριναν παρόμοιες μεθόδους, μέρος των οποίων ήταν και οι μέθοδοι LSI, pLSI και LDA. Έδειξαν, ότι η μέθοδος LSI υπολείπεται της ικανότητας της μοντελοποίησης "μίγματος", δηλαδή της ανάθεσης πιθανοτικών "μιγμάτων" των θεμάτων και της ικανότητας να γενικεύει, η μέθοδος pLSI υπολείπεται της ικανότητας της γενίκευσης, ενώ ο LDA δεν υστερεί σε τίποτα από τα παραπάνω. Βέβαια, όλες αυτές οι μέθοδοι, όπως και η μέθοδος Term Frequency–Inverse Document Frequency (TF-IDF) που ήταν μέρος της μελέτης τους, έχουν την ικανότητα να μειώνουν τις διαστάσεις των δεδομένων. Ανέφεραν επίσης, ότι η μέθοδος TF-IDF, υπολείπεται των προαναφερόμενων ικανοτήτων, αλλά και της ικανότητας των θεμάτων να μαθαίνουν από την ανάλυση των αρχικών κειμένων (σημασιολογικός σχολιασμός). Τονίζουν ότι κύριο μειονέκτημα της μεθόδου LSI, είναι η έλλειψη προσαρμογής ενός μοντέλου στα δεδομένα για την αναπαράσταση των εγγράφων σε θέματα. Η πιθανοτική μέθοδος pLSI, αν και είναι ικανή να αναθέτει πολλαπλά θέματα σ' ένα έγγραφο, δεν μπορεί να γενικεύσει σε έγγραφα που δεν αποτέλεσαν μέρος της εκπαίδευσης, όπως έδειξε και η προηγούμενη έρευνα αλλά και ότι η μέθοδος pLSI, είναι ευαίσθητη στην υπερ-προσαρμογή, καθώς ο αριθμός των παραμέτρων αυξάνεται γραμμικά με τον αριθμό των εγγράφων που έχουμε. Αυτό αποτελεί σοβαρό πρόβλημα για ένα μοντέλο. Έτσι, επικεντρωθήκαμε στην ανάπτυξη ενός LDA μοντέλου για την ταξινόμηση των εγγράφων, τροποποιώντας τις παραμέτρους του για την αποτελεσματικότερη απόδοση στην εξαγωγή θεμάτων, βασιζόμενο στα επιστημονομετρικά δεδομένα που συγκεντρώσαμε.

Σχετικά με τα πειράματα συσταδοποίησης εγγράφων, μία έρευνα [48] των Bapuji Rao και Brojo Kishore Mishra, αποτέλεσε το κίνητρο για τη συσταδοποίηση εγγράφων που "μοιράζονται" παρόμοια θέματα. Συγκεκριμένα, αυτή η έρευνα εισάγει μία νέα προσέγγιση για την ομαδοποίηση εγγράφων κειμένου με βάση ένα σύνολο λέξεων, χρησιμοποιώντας τεχνικές εξόρυξης γραφημά-

των. Η σχέση εγγράφου-λέξης, μπορεί να αναπαρασταθεί ως ένα διμερές γράφημα που φαίνεται και στο Σχήμα (Σχ.) 2.1, ενώ ολόκληρη η συσταδοποίηση, παρουσιάζεται ως δευτερεύοντα γραφήματα. Αρχικά, κατασκεύασαν τον πίνακα εγγράφου-λέξης, όπου η κάθε καταχώρηση αντιπροσωπεύει τον αριθμό εμφανίσεων του κάθε όρου (λέξης) στο αντίστοιχο έγγραφο. Αυτός ο πίνακας, αποτέλεσε και κομμάτι του δικού μας πειράματος. Τον αριθμό των εγγράφων τον συμβολίζουν με *n* και τον αριθμό των λέξεων με *m*. Χρησιμοποιώντας τον παραπάνω πίνακα συχνοτήτων, ο αλγόριθμος τους σχημάτισε *n* αριθμό συστάδων από έγγραφα, για τον ίδιο αριθμό λέξεων. Έτσι, η κάθε συστάδα αποτελείται από *m* ή και λιγότερα έγγραφα. Η σχέση μεταξύ εγγράφων και λέξεων, συμβολίζεται με έναν αριθμό πάνω σε μία ακμή που ενώνει το έγγραφο με την κάθε λέξη που εμφανίζεται μέσα σ' αυτό. Για παράδειγμα, αν υπάρχει στην ακμή ο αριθμός 10, δηλώνει ότι σ' εκείνο το έγγραφο, η λέξη που προσπίπτει η ακμή, εμφανίζεται 10 φορές μέσα σ' αυτό.



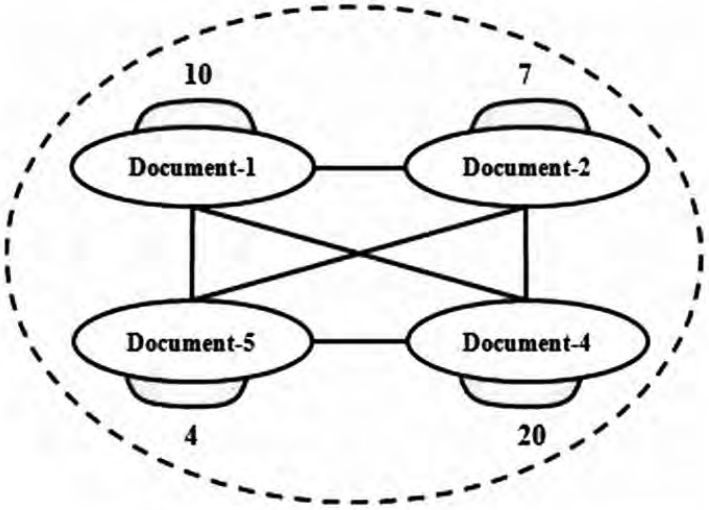
Σχήμα 2.1: Διμερές γράφημα εγγράφου-λέξης (Πηγή: [48])

Όλη αυτή η σχέση μπορεί να αναπαρασταθεί σ' έναν σταθμισμένο πίνακα πρόσπτωσης. Αυτός ο πίνακας προτάθηκε στο άρθρο [47] και παρουσιάζεται στο Σχ. 2.2. Αυτό το άρθρο, το έγραψαν οι Bapuji Rao και Mitra Anirban και επικεντρώνεται στη δημιουργία τεχνικών γραφημάτων εξόρυξης. Είχε ως απώτερο σκοπό την ανίχνευση κοινότητας στα πλαίσια του κοινωνικού δικτύου με την ανάπτυξη δικού τους αλγορίθμου, δηλαδή στην ανάλυση σχέσεων μεταξύ κοινωνικών δικτύων. Έτσι, οι γραμμές του πίνακα εκπροσωπούν τα έγγραφα (*m*), ενώ οι στήλες του εκπροσωπούν τις λέξεις (*n*). Με αυτόν τον τρόπο, σχημάτισαν *n* συστάδες, όπου η κάθε συστάδα είναι η λέξη και θεωρείται ως υπο-γράφημα εγγράφων για μία συγκεκριμένη λέξη, ενώ τα μέλη της είναι τα έγγραφα που περιέχουν έστω και μία φορά τον αντίστοιχο όρο. Παράδειγμα συσταδοποίησης για μία λέξη, δίνεται στο Σχ. 2.3. Η προαναφερόμενη έρευνα, μας έδωσε το κίνητρο για τη συσταδοποίηση εγγράφων που "μοιράζονται" παρόμοια θέματα, χρησιμοποιώντας τον αλγόριθμο K-Means, όπου η κάθε συστάδα αποτελεί το θέμα, ενώ τα μέλη της, είναι έγγραφα που με βάση τον πιθανοτικό πίνακα εγγράφου-θέματος που αναπτύξαμε, διαθέτουν και το μεγαλύτερο πιθανοτικό σκορ σχετικά με το κάθε θέμα.

Η συσταδοποίηση εγγράφων, με συντεταγμένες το μέγεθος των τίτλων τους και τον αριθμό των συνολικών τους αναφορών την περίοδο 1900-2009, στηρίχθηκε σε μία έρευνα [34] των Adrian Letchford, Helen Susannah Moat και Preis Tobias. Αυτή η έρευνα βασίζεται σ' ένα δείγμα 140.000 εγγράφων, για να διερευνηθεί εάν το μήκος του τίτλου ενός εγγράφου, έχει οποιαδήποτε σχέση με τον αριθμό των παραπομπών που λαμβάνει. Η ανάλυσή τους, παρέχει αποδεικτικά στοιχεία, τα οποία αποδεικνύουν ότι τα περιοδικά που δημοσιεύουν έγγραφα με τίτλους μικρότερου μήκους,

	Word-1	Word-2	Word-3	Word-4	.....	Word-n
Document-1	10	7	0	0	.....	5
Document-2	7	3	6	25	.....	0
Document-3	0	4	25	0	.....	0
Document-4	20	0	4	0	.....	9
Document-5	4	0	7	9	.....	10
.....	.....	.....	.....	.....	.....	.....
Document-m	0	0	0	50	.....	5

Σχήμα 2.2: Σταθμισμένος πίνακας πρόσπτωσης εγγράφου-όρου (Πηγή: [48])



Σχήμα 2.3: Συσταδοποίηση της γραφικής παράστασης των εγγράφων για μία λέξη (Πηγή: [48])

λαμβάνουν και τις περισσότερες αναφορές ανά έγγραφο. Ανακάλυψαν ότι από τα 20.000 πιο κορυφαία δημοσιευμένα έγγραφα το έτος 2010, εκείνα με τα μικρότερη μήκη τίτλων, ήταν αυτά που έλαβαν τις περισσότερες παραπομπές. Έτσι, εμπνευστήκαμε από αυτήν την έρευνα και προσπαθήσαμε να ανακαλύψουμε τη σχέση μήκους τίτλων και αναφορών, πραγματοποιώντας μία ανάλογη συσταδοποίηση. Έχει ως απώτερο σκοπό την ανάδειξη ύπαρξης πολλών αστάθμητων παραγόντων που μπορούν να συντελέσουν στην αναφορά ενός εγγράφου, χωρίς να δίνεται μονομερή σημασία στον παράγοντα του μήκους των τίτλων των εγγράφων, ο οποίος εξακολουθεί να δημιουργεί ερωτήματα, για το αν σχετίζεται άμεσα με το πλήθος των αναφορών που λαμβάνουν τα έγγραφα.

Για το τελευταίο κομμάτι των πειραμάτων που αφορά την πρόβλεψη μελλοντικών αναφορών που θα λάβει ένα έγγραφο την επόμενη τριετία, εμπνευστήκαμε από ένα άρθρο [4] των Ali Abrishami και Sadegh Aliakbary. Συγκεκριμένα, στην προαναφερόμενη εργασία, πρότειναν μία μέθοδο για την πρόβλεψη μακροπρόθεσμων παραπομπών ενός εγγράφου (π.χ από τον 5^o έως τον 15^o χρόνο μετά τη δημοσίευση), με βάση τον αριθμό των αναφορών του κατά τα πρώτα έτη μετά τη δημοσίευση (από 3 έως 5 έτη). Για να πραγματοποιήσουν την εκπαίδευση του μοντέλου που θα πετύχαινε κάτι τέτοιο, χρησιμοποίησαν όπως και εμείς, τεχνητά νευρωνικά δίκτα. Δεν λήφθηκαν

υπ' όψιν άλλες πηγές πληροφοριών όπως ο συγγραφέας, η απήχηση ενός περιοδικού, το περιεχόμενο του κ.λπ.

Κατασκεύασαν ένα Recurrent Neural Network που αποτελείται από δύο ανεξάρτητα νευρωνικά δίκτυα, τα οποία ονομάζονται κωδικοποιητής και αποκωδικοποιητής. Οι είσοδοι τροφοδοτούνται στον κωδικοποιητή και οι έξοδοι λαμβάνονται από τον αποκωδικοποιητή. Στον τομέα του αποκωδικοποιητή, χρησιμοποιείται ένα πυκνό στρώμα, όπως και στο δικό μας το μοντέλο, για να παράγει την έξοδο (προβλεπόμενος αριθμός αναφορών). Η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε σ' όλα τα επίπεδα του νευρωνικού τους δικτύου, ήταν η ReLU. Η διάρκεια σε περιόδους χρόνου εκπαίδευσης του δικού τους νευρωνικού, ήταν 100. Το μέγεθος της παρτίδας των δεδομένων ήταν 256 έγγραφα, ενώ αντί για τον αλγόριθμο Adam που εφαρμόσαμε εμείς, χρησιμοποίησαν ως μέθοδο βελτιστοποίησης τον RMSProp, με ρυθμό μάθησης 0.00001.

Ως μέτρα αξιολόγησης του μοντέλου τους, χρησιμοποίησαν τη συνάρτηση απώλειας RMSE και την R^2. Η συνάρτηση RMSE και η R^2, συναντώνται και στο δικό μας μοντέλο και θα τις αναλύσουμε στο Κεφάλαιο 5. Στα πειράματα τους, αναφορικά με τους διαφορετικούς αριθμούς αναφορών που λαμβάνουν ως ακολουθία εισόδου για το νευρωνικό, μετά τη χρονιά δημοσίεσης των εγγράφων, έδειξαν ότι και οι δύο συναρτήσεις που εφαρμόστηκαν για την αξιολόγηση του μοντέλου, επέδειξαν ακριβέστερα αποτελέσματα όταν εισέρχεται στο μοντέλο περισσότερη πληροφορία σχετικά με το πλήθος των αναφορών. Αυτός ήταν και ο λόγος που η δική μας μέθοδος στηρίχθηκε σε Recurrent Neural Network, με τη διαφορά ότι ο κωδικοποιητής και ο αποκωδικοποιητής είναι τύπου Long Short Term Memory και λαμβάνει ως είσοδο 12 χρονιές που συμβολίζουν το συνολικό αριθμό παραπομπών των εγγράφων μας, μετά τη χρονιά δημοσίευσης αυτών.

2.2 Τεχνητή Νοημοσύνη

2.2.1 Εισαγωγή

Η T.N αναφέρεται στον κλάδο της επιστήμης των υπολογιστών, ο οποίος ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων. Τα τελευταία μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς, τα οποία υπονοούν έστω και στοιχειώδη ευφυΐα, όπως είναι η μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα, επίλυση προβλημάτων κ.ο.κ. Εξελίσσεται τις τελευταίες δεκαετίες με ραγδαίο ρυθμό. Αποτελεί το σταυροδρόμι πολλών πεδίων, όπως της επιστήμης των υπολογιστών ή της ψυχολογίας και κύριος στόχος της είναι, να αναπαράγει την ανθρώπινη ευφυΐα μέσα από δικά της μοντέλα. Πολλές φορές είναι αδύνατο να δώσουμε ακριβή ορισμό της νοημοσύνης, λόγω των πολλών παραμέτρων που τη χαρακτηρίζουν. Ο όρος της T.N επινοήθηκε το 1956, αλλά έχει γίνει πιο δημοφιλές σήμερα λόγω του αυξημένου όγκου δεδομένων, των προηγμένων αλγορίθμων και των βελτιώσεων, στην ισχύ των υπολογιστών και την αποθήκευση των δεδομένων.

2.2.2 Ιστορική Αναδρομή

Τη δεκαετία του 1940, ο McCulloch και ο Pitts, πρότειναν ένα μοντέλο τεχνητών νευρώνων που είχε τη δυνατότητα να μαθαίνει και να υπολογίζει συναρτήσεις. Αυτό αποτέλεσε και την πρώτη

μαθηματική περιγραφή τεχνητού νευρωνικού δικτύου. Το 1950, ο μαθηματικός Alan Turing, ο οποίος θεωρείται ο «πατέρας της επιστήμης των υπολογιστών», πρότεινε τη δοκιμή Τούρινγκ, από την οποία ήθελε να εξακριβώσει αν μία μηχανή διαθέτει γνωστικές ικανότητες και τη δυνατότητα να σκεφτεί. Ο όρος Τ.Ν, δόθηκε επίσημα το 1956 από τον John McCarthy, όταν πραγματοποιήθηκε η πρώτη ακαδημαϊκή διάσκεψη από Αμερικανούς επιστήμονες σχετικά με το θέμα. Τη χρονιά αυτή, παρουσιάστηκε για πρώτη φορά το Logic Theorist, το οποίο ήταν το πρώτο πρόγραμμα της Τ.Ν που σχεδιάστηκε σκόπιμα για να μιμηθεί τις ικανότητες επίλυσης προβλημάτων ενός ανθρώπου. Δύο χρόνια αργότερα, το 1958, ο ίδιος ο McCarthy σχεδιάζει την LISP, την πρώτη οικογένεια συναρτησιακών γλωσσών προγραμματισμού. Τα επόμενα χρόνια, κάνουν την εμφάνιση τους οι γενετικοί αλγόριθμοι, καθώς και ένα πιο προχωρημένο νευρωνικό δίκτυο, το perceptron, από τον Rosenblatt. Το 1970, έχουμε τη δημιουργία της Prolog, μίας γλώσσας λογικού προγραμματισμού και ακολουθεί το 1980, όπου έχουμε την ανάπτυξη πολυεπίπεδου perceptron. Το πολυεπίπεδο perceptron διαθέτει εισόδους, κρυφά επίπεδα όπου ορίζεται ο αριθμός των νευρώνων, καθώς και εξόδους, όπως φαίνεται και στο Σχ. 2.4.

Τα επόμενα χρόνια, η Τ.Ν αναπτύσεται ακόμα περισσότερο λόγω της μεγάλης ανάπτυξης του διαδικτύου, με την εμφάνιση των (ευφυών) πρακτόρων και της απήχησής τους, στο τομέα της επικοινωνίας μεταξύ χρήστη και λογισμικού στο ψηφιακό κόσμο. Αξιοσημείωτο γεγονός αποτελεί η νίκη του υπολογιστή Deep Blue της IBM έναντι του τότε παγκόσμιου πρωταθλητή στο σκάκι Garry Kasparov που οδήγησε πολλούς ανθρώπους να αναρωτηθούν ποια είναι τα όρια της Τ.Ν. Στα τέλη της δεκαετίας του ’90, δημιουργείται το πρώτο σκυλί AIBO της Sony που είναι ένα από τα πρώτα αυτόνομα κατοικίδια με δυνατότητα κίνησης στο χώρο και εκδήλωσης συναισθημάτων. Έπειτα, η Apple προωθεί το iPhone 4S με την προσωπική βοηθό Siri που βοηθάει τον χρήστη να πραγματοποιήσει δικές του επιθυμίες, δίνοντας φωνητικές εντολές.

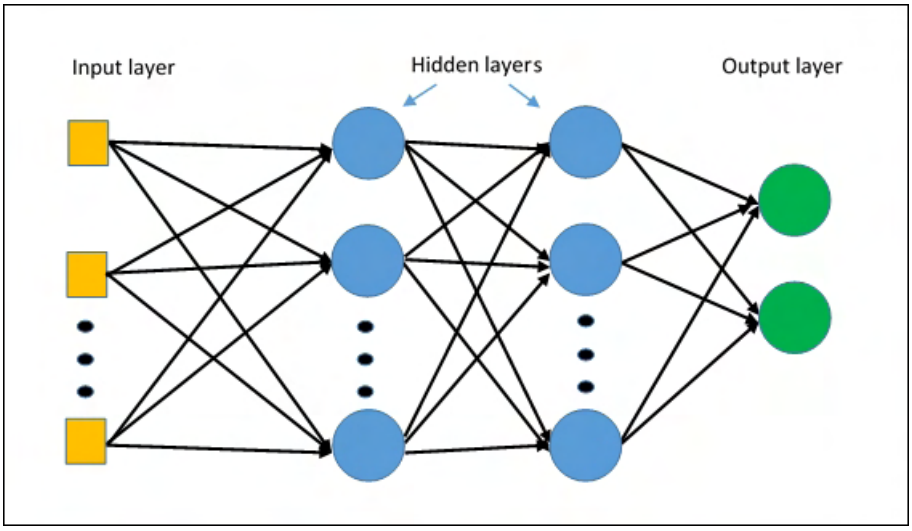
Σήμερα, η σύγχρονη Τ.Ν αποτελεί ένα από τα πλέον μαθηματικοποιημένα πεδία της πληροφορικής που μπορεί να διεκπεραιώσει πλήθος διεργασιών παρέχοντας παράλληλα, με ανθρωπόμορφες αλληλεπιδράσεις, υποστήριξη στη λήψη αποφάσεων. Αναμφισβήτητα, η πρόδος είναι εντυπωσιακή και θα λέγαμε και αναμενόμενη, με πολλαπλά οφέλη σε διάφορα επίπεδα, ωστόσο η Τ.Ν δεν πρέπει να αποτελέσει στο μέλλον υποκατάστατο των ανθρώπων.

2.2.3 Ερευνητικοί Χώροι της Τεχνητής Νοημοσύνης

Η Τ.Ν είναι ένα πεδίο μελέτης που σχετίζεται με πολλές θεωρίες και τεχνολογίες, σε αντίθεση με παλαιότερες πεποιθήσεις που την ήθελαν να είναι μόνο ρομπότ με ανθρώπινα χαρακτηριστικά. Παρακάτω, θα δούμε ορισμένους ερευνητικούς χώρους στους οποίους συναντάται.

- **Μηχανική Μάθηση**. Ο τομέας αυτός της Τ.Ν, ασχολείται με τη μελέτη αλγορίθμων που βελτιώνουν την απόδοση ενός συστήματος κατά την εκτέλεση μίας συγκεκριμένης εργασίας, ώστε να μπορεί το σύστημα να αξιοποιεί προηγούμενη γνώση και εμπειρία, χωρίς δηλαδή να καθίσταται ανάγκη να προγραμματιστεί εκ νέου.

- **Νευρωνικά Δίκτυα**. Είναι ένα σύνολο αλγορίθμων, εμπνευσμένοι από τη λειτουργία του ανθρώπινου εγκεφάλου, οι οποίοι έχουν σχεδιαστεί για να αναγνωρίζουν πρότυπα. Ονομά-



Σχήμα 2.4: Παράδειγμα ενός Multilayer Perceptron (Πηγή: [41])

ζονται αλλιώς και τεχνητά νευρωνικά δίκτυα. Αποτελούνται από ένα στρώμα εισόδου, ένα στρώμα εξόδου και ένα ή περισσότερα κρυμμένα στρώματα.

- **Βαθιά Μάθηση**. Είναι υποπεδίο της Μ.Μ και διαθέτει ισχυρές τεχνικές που αφορούν την εκπαίδευση στα νευρωνικά δίκτυα.

- **Επίλυση προβλημάτων**. Αντικείμενο του τομέα είναι η μελέτη ευφυών αλγορίθμων εύρεσης λύσεων.

- **Ρομποτική**. Ένας σύγχρονος τομέας της αυτοματοποίησης, με αντικείμενο τη μελέτη, το σχεδιασμό και τη λειτουργία ρομποτικών συστημάτων, σε συνδυασμό με ηλεκτρομηχανικές διατάξεις.

2.3 Μηχανική Μάθηση

2.3.1 Εισαγωγή

Στις μέρες μας, λόγω της μεγάλης ποσότητας δεδομένων (κείμενα, εικόνες, βίντεο κ.ο.κ), καθίσταται η ανάγκη να αναλύσουμε και να επεξεργαστούμε αυτά τα δεδομένα, με απώτερο σκοπό την εξαγωγή χρήσιμων πληροφοριών που αρχικά δεν είναι εμφανής. Αυτό μπορεί να πραγματοποιηθεί μέσω της εξόρυξης δεδομένων και της Μ.Μ. Αποτελεί αναπόσπαστο κομμάτι της Τ.Ν και ασχολείται όπως προηγουμένως αναφέραμε, με το σχεδιασμό αλγορίθμων, ώστε ο υπολογιστής να μαθαίνει αξιοποιώντας την εμπειρία από προηγούμενα προβλήματα και να μην χρειάζεται επαναπρογραμματισμό. Στηρίζεται στη δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων και αποτελεί έναν από τους ταχύτερα αναπτυσσόμενους τομείς της επιστήμης των υπολογιστών, με εφαρμογές σε πολλά πεδία, κάποια από τα οποία εμφανίζονται στο Σχ. 2.5.

Σχήμα 2.5: Πεδία της Μ.Μ (Πηγή: [36])

,

2.3.2 Η Διαδικασία Εξαγωγής Μοντέλου

Η διαδικασία με την οποία μπορεί να επιλεχθεί το καταλληλότερο μοντέλο δεν είναι απλή, γιατί υπάρχει μία ποικιλία αλγορίθμων που επαναληπτικά, μαθαίνουν από τα δεδομένα για να μπορούν να βελτιώνονται. Η διαδικασία ξεκινά με τη συλλογή και την προετοιμασία των δεδομένων που απαιτεί και το περισσότερο χρόνο. Τα δεδομένα αφου περάσουν από αυτό το στάδιο, διαχωρίζονται σε δεδομένα εκπαίδευσης, για την εκπαίδευση του μοντέλου και σε δεδομένα δοκιμής, για την αξιολόγηση της αποτελεσματικότητάς του. Τα δεδομένα εκπαίδευσης αποτελούν την είσοδο του μοντέλου. Εν συνεχεία, με βάση τα επιθυμητά αποτελέσματα που έχουμε προκαθορίσει, γίνεται η λεγόμενη εκπαίδευση μοντέλου. Τέλος, πραγματοποιείται η εκτίμηση των αποτελεσμάτων με βάση τα δεδομένα δοκιμής και σε συνδυασμό με τη ρύθμιση των υπερ-παραμέτρων του μοντέλου, όπως είναι ο ρυθμός μάθησης, ο αριθμός επαναλήψεων κ.λπ, εξάγεται το τελικό μοντέλο.

2.3.3 Κατηγορίες της Μηχανικής Μάθησης

Υπάρχουν κάποιες παραλλαγές στον τρόπο κατηγοριοποίησης των τύπων των αλγορίθμων της Μ.Μ, αλλά συνήθως, μπορούν να χωριστούν σε κατηγορίες ανάλογα με το σκοπό τους και οι κύριες κατηγορίες είναι οι εξής:

- **Εποπτευόμενη Μάθηση**. Αποτελεί την πιο συνηθισμένη μέθοδο. Σε αυτήν την κατηγορία πραγματοποιείται εκπαίδευση του αλγορίθμου, με βάση τις εισόδους αλλά και τις εξόδους που έχουμε προσδιορίσει αρχικά. Η αποτελεσματικότητα της διαδικασίας στηρίζεται στην καλύτερη περιγραφή των δεδομένων εισόδου, δηλαδή εκείνης που για μία συγκεκριμένη είσοδο, προβλέπει την καταλληλότερη για εμάς έξοδο. Η οπισθοδρόμηση και η ταξινόμηση είναι τα κυριότερα παραδείγματα αυτής της κατηγορίας.

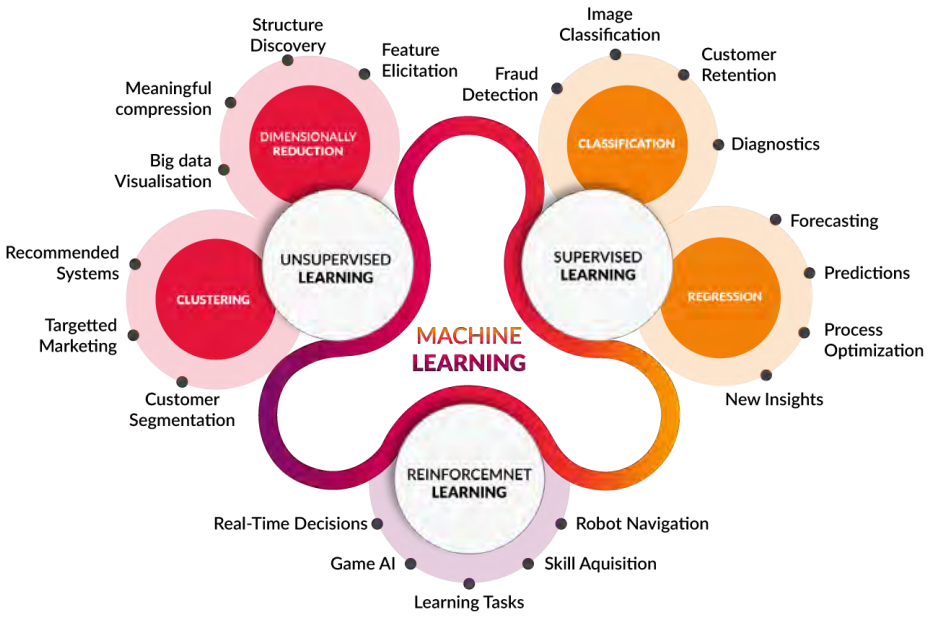
- **Αυτοελεγχόμενη Μάθηση**. Σε αυτήν την υποπερίπτωση της εποπτευόμενης μάθησης, τα

μοντέλα μαθαίνουν εντελώς από μόνα τους, χωρίς τη γνώση για επιθυμητά αποτελέσματα, όπου τα τελευταία ορίζονται από τα δεδομένα εισόδου με χρήση ευρετικών αλγορίθμων.

- **Μη εποπτευόμενη μάθηση**. Η εκπαίδευση του αλγορίθμου γίνεται με βάση μόνο τα δεδομένα εισόδου. Δεν υπάρχουν επιθυμητά δεδομένα εξόδου και τα τελικά αποτελέσματα στηρίζονται στην εξαγωγή προτύπων, μετά την εκπαίδευση του μοντέλου. Αυτοί οι αλγόριθμοι, είναι ιδιαίτερα χρήσιμοι στις περιπτώσεις που ο χρήστης δεν γνωρίζει τι αναζητάει από τα δεδομένα. Οι κύριοι τύποι της μη εποπτευόμενης μαθήσης είναι η μείωση των διαστάσεων και η συσταδοποίηση.

- **Ημι-εποπτευόμενη Μάθηση**. Στη συγκεκριμένη κατηγορία που θα μπορούσαμε να πούμε ότι είναι μεταξύ εποπτευόμενης και μη, μάθησης, για κάποια δείγματα επιθυμούμε μία συγκεκριμένη έξοδο, ενώ για κάποια άλλα δεν γνωρίζουμε ποια θα ήταν η καταλληλότερη έξοδος.

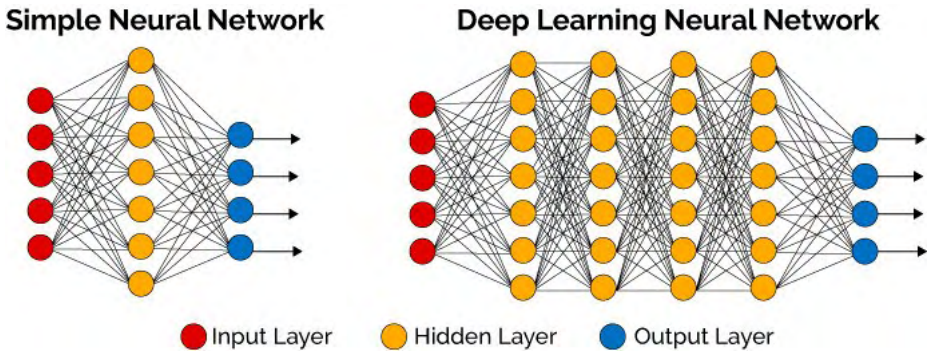
- **Ενισχυμένη Μάθηση**. Είναι ένας τύπος Μ.Μ που δίνει τη δυνατότητα στις μηχανές να προσδιορίζουν αυτόματα την ιδανική συμπεριφορά, προκειμένου να μεγιστοποιηθεί η απόδοση του μοντέλου. Στην παρούσα κατηγορία, οι αλγόριθμοι μαθαίνουν συνεχώς από το περιβάλλον, με επαναληπτικό τρόπο. Οι μηχανές εκπαιδεύονται με τη μέθοδο δοκιμής και λάθους και με βάση την πλήρη διερεύνηση του φάσματος όλων των πιθανών καταστάσεων, προσπαθούν να λάβουν ακριβείς αποφάσεις. Στο Σχ. 2.6, εμφανίζεται η σχηματική κατηγοριοποίηση των βασικών κατηγοριών της Μ.Μ, όπως τις αναλύσαμε προηγουμένως.



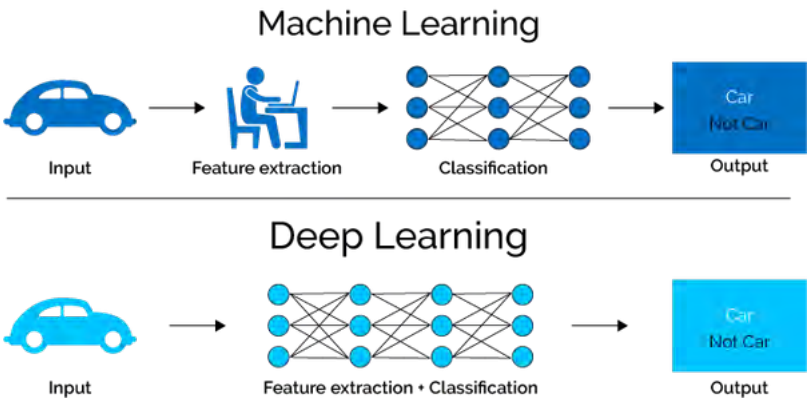
Σχήμα 2.6: Κατηγορίες της Μ.Μ (Πηγή: [14])

2.4 Βαθιά Μάθηση

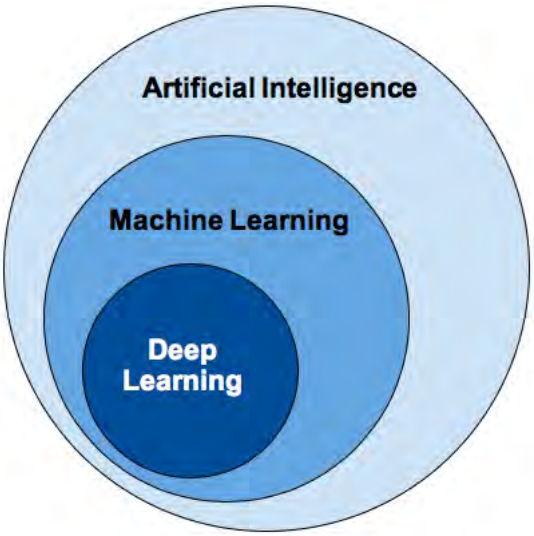
Η Βαθιά Μάθηση (Β.Μ) είναι ένα υποπεδίο της Μ.Μ που ασχολείται με αλγορίθμους εμπνευσμένους από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου και αφορά τα τεχνητά νευρωνικά δίκτυα. Αναλύει δεδομένα και μαθαίνει χαρακτηριστικά απευθείας απ' αυτά. Παρουσιάζει μεγαλύτερη απόδοση από παλαιότερους αλγορίθμους μάθησης και αυτό συμβαίνει λόγω των πολυεπίπεδων στρωμάτων που χρησιμοποιούνται στην ανάπτυξη ενός μοντέλου. Αυτά τα επίπεδα στρωμάτων, μπορεί να είναι μερικές δεκάδες και να φτάσουν μέχρι εκατοντάδες, έχοντας την ικανότητα να εκπαιδεύονται μόνα τους από τα δεδομένα εκπαίδευσης που αναφέραμε στην Υποενότητα 2.3.2. Ο όρος "βαθιά", συνήθως αναφέρεται στον αριθμό των κρυμμένων στρωμάτων μέσα στο νευρωνικό δίκτυο που μπορεί να φτάσει μέχρι 150, σε αντίθεση με ένα παραδοσιακό νευρωνικό που αποτελείται από 1 έως 2 με 3 κρυμμένα στρώματα, όπως απεικονίζεται και στο Σχ. 2.7. Η αρχιτεκτονική της Β.Μ, είναι και η κυριότερη διαφορά με τη Μ.Μ, καθώς η τελευταία αποτελείται από ένα επίπεδο στρωμάτων όπως φαίνεται και στο Σχ. 2.8, ενώ η πρώτη, μπορεί να παρουσιάσει υψηλή πολυπλοκότητα με πολούς νευρώνες. Κλείνοντας την ενότητα αυτή, στο Σχ. 2.9, απεικονίζεται η αναπαράσταση της σχέσης μεταξύ της Τ.Ν, της Μ.Μ και της Β.Μ, με απώτερο σκοπό τη σωστή κατανόηση και ιεράρχηση των προαναφερθέντων εννοιών. Παρατηρούμε ότι η Μ.Μ, είναι υποπεδίο της Τ.Ν, ενώ η Β.Μ, αποτελεί υποπεδίο της Μ.Μ.



Σχήμα 2.7: Απλό νευρωνικό δίκτυο έναντι ενός πολύπλοκου νευρωνικού δικτύου (Πηγή: [19])



Σχήμα 2.8: Μ.Μ και Β.Μ (Πηγή: [63])



Σχήμα 2.9: Τ.Ν, Μ.Μ και Β.Μ (Πηγή: [7])

2.5 Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων, αποτελεί μία από τις νεοφερμένες μεθόδους που χρησιμοποιούν οι εταιρείας έρευνας αγοράς. Είναι η διαδικασία εξεύρεσης ανωμαλιών, μοτίβων και συσχετισμών, μέσα σε μεγάλα σύνολα δεδομένων για την πρόβλεψη των αποτελεσμάτων. Η Μ.Μ χρησιμοποιεί τις τεχνικές της εξόρυξης, για τη δημιουργία μοντέλων σχετικά με το τι συμβαίνει πίσω από ορισμένα δεδομένα, ώστε να μπορεί να προβλέψει μελλοντικά αποτελέσματα. Ο πρωταρχικός στόχος της εξόρυξης δεδομένων, είναι η πρόβλεψη. Είναι παρόμοια με τη Μ.Μ και επικεντρώνεται περισσότερο στις πρακτικές πτυχές της ανάπτυξης αλγορίθμων και στην πλαισίωση διαφόρων υποθέσεων και όχι στην απόδειξη αυτών. Προσπαθεί να δημιουργήσει τη διαίσθηση για το τι πραγματικά συμβαίνει σε συγκεκριμένα δεδομένα, χρησιμοποιώντας κυρίως στο μαθηματικό κομμάτι, τον κλάδο της στατιστικής, συνδυάζοντας το τελευταίο με προγραμματιστικό περιβάλλον. Τα δεδομένα εξόρυξης, μαζί με τα πρότυπα και τις υποθέσεις, μπορούν να χρησιμοποιηθούν ως βάση, τόσο για την Τ.Ν όσο και για τη Μ.Μ.

Κεφάλαιο 3

Ο Τομέας της Επιστημονομετρίας

Σ' αυτό το κεφάλαιο θα κάνουμε μία εισαγωγή στον τομέα της επιστημονομετρίας και σε έννοιες που είναι άρρητες συνδεδεμένες με αυτόν τον κλάδο. Θα οριστούν μετρικές επιστήμες, παρόμοιες με την επιστημονομετρία που είναι απαραίτητες για τη γενικότερη κατανόηση των πειραμάτων και συνυφασμένες με το αντικείμενο της παρούσας εργασίας. Τέλος, αποσαφηνίζονται οι παράγοντες που συντελούν στην αναγνωρισιμότητα και στην αναφορά των εγγράφων, τόσο στην επιστημονική όσο και στην ευρεία κοινότητα.

3.1 Εισαγωγή

Η έννοια της επιστημονομετρίας καθορίστηκε για πρώτη φορά το 1971, από τον V.V. Nalimov, ως την «ανάπτυξη των ποσοτικών μεθόδων της έρευνας για την ανάπτυξη της επιστήμης ως πληροφοριακής διαδικασίας». Κύριος στόχος της επιστημονομετρίας, είναι η εύρεση του αντίκτυπου που έχουν οι συγγραφείς, τα άρθρα, τα περιοδικά κ.ο.κ. Επιδιώκει επίσης, να κατανοήσει τη συμπεριφορά των επιστημονικών αναφορών και να δημιουργήσει δείκτες που συμβάλουν στην αξιολόγηση της απόδοσης και της παραγωγικότητας. Επιπρόσθετα, συμβάλει στον καθορισμό του επιπέδου της επιστημονικής αλλά και της τεχνολογικής ανάπτυξης. Επικεντρώνεται στην επικοινωνία μεταξύ των επιστημών, ειδικότερα των κοινωνικών και των ανθρωπιστικών επιστημών.

Τα τελευταία χρόνια, ο κλάδος της επιστημονομετρίας έχει διαδραματίσει σημαντικό ρόλο στην καταγραφή και στην εκτίμηση της δυναμικής της επιστήμης. Η δυνατότητα να καταγράφει και να αναλύει επιστημονικά δεδομένα, είναι και ο λόγος που συνέβαλε στην εξαιρετικά γρήγορη ανάπτυξή της. Ο όγκος της δημοσιευμένης βιβλιογραφίας έχει αυξηθεί ραγδαία και το ποσοστό των ερευνητών έχει μεγαλώσει αρκετά, δημιουργώντας ένα ανταγωνιστικό περιβάλλον. Δεδομένων των συνθηκών και λόγω της αύξησης της επιστημονικής πολυμορφίας, καθίσταται η ανάγκη για ανάλυση αυτών των δεδομένων, επεξεργασία και εξαγωγή συμπερασμάτων, με απώτερο σκοπό τον εντοπισμό των αναδυόμενων τάσεων και την αντιμετώπιση των νέων προκλήσεων.

3.2 Ορισμός Συναφών Μετρικών Επιστημών

Ο τομέας της βιβλιοθήκης και της επιστήμης των πληροφοριών, αποτελείται από τα υποπεδία "Informetrics", "Bibliometrics", "Scientometrics", "Cybermetrics", και "Webometrics". Τα τελευταία 2 υποπεδία, είναι σχετικά καινούρια και αναπτύχθηκαν με τη βοήθεια της ανάπτυξης του διαδικτύου και αφορούν τα ηλεκτρονικά μέσα επικοινωνίας. Τα κυριότερα υποπεδία είναι τα 3 πρώτα και συχνά ο ακαδημαϊκός κόσμος, τα αποκαλεί τα 3 μέτρα. Αυτές οι ορολογίες, αναπτύχθηκαν στους τομείς της επιστήμης των πληροφοριών, της βιβλιοθήκης, της φιλολογίας και της επιστήμης.

Τα παραπάνω συναφή μετρικά υποπεδία, παρόλο που "μοιράζονται" κοινές αρχές και μεθόδους, έχουν διαφορετική περιοχή μελέτης. Στην καθημερινότητα, δεν είναι λίγες οι φορές που παρατηρούμε να υπάρχει μία αλληλοεπικάλυψη μεταξύ αυτών των γειτονικών υποπεδίων. Μάλιστα, ο Wormell το 1998, περιέγραψε τη χαοτική κατάσταση αυτών των ορολογιών, αναφέροντας ότι «Οι ατομικές ταυτότητες των υποπεδίων "Bibliometrics", "Informetrics", "Scientometrics" και "Technometrics", δυστυχώς δεν είναι πολύ σαφείς και υπάρχει χάος στην ορολογία». Βέβαια, έχουν αναπτυχθεί σε διαφορετικούς βαθμούς με το πέρασμα των χρόνων και είναι αναγνωρίσιμα από τον ακαδημαϊκό κόσμο. Παρακάτω, θα δούμε αυτές τις μετρικές επιστήμες που κύριος στόχος τους είναι η καθιέρωση ποσοτικοποιημένων μεθόδων ανάλυσης και μετρήσεων, για τις επιστημονικές εργασίες.

- **Informetrics**. Είναι ένας ευρύτερος όρος που μελετά τις ποσοτικές πτυχές της πληροφορίας και δεν πρέπει να συγχέεται με τον τομέα της πληροφορικής. Περιλαμβάνει την ηλεκτρονική επικοινωνία των μέσων μαζικής ενημέρωσης, συμπεριλαμβανομένου του διαδικτύου και του παγκόσμιου ιστού, των βιβλίων και των περιοδικών.

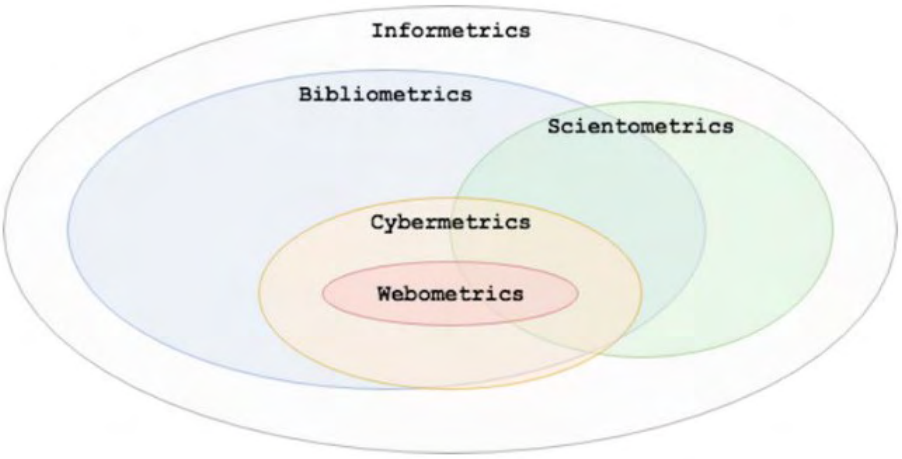
- **Bibliometrics**. Το συγκεκριμένο υποπεδίο, μελετά τις ποσοτικές πτυχές των καταγεγραμμένων πληροφοριών των βιβλίων και των δημοσιεύσεων γενικότερα. Μπορεί να ταξινομηθεί σε 2 υποκατηγορίες. Η μία κατηγορία είναι υπεύθυνη για την περιγραφή χαρακτηριστικών μίας βιβλιογραφίας (περιγραφικές μελέτες) και η άλλη εξετάζει τις σχέσεις που αναπτύσσονται μεταξύ των συνιστωσών μίας λογοτεχνίας (συμπεριφορικές μελέτες).

- **Scientometrics**. Ο τομέας της επιστημονομετρίας, ασχολείται με την ποσοτική μελέτη διαφόρων ειδών διεργασιών νοημοσύνης, στην ανάπτυξη της επιστήμης. Συγκεκριμένα, χρησιμοποιεί μαθηματικές μεθόδους για να ποσοτικοποιήσει το επιστημονικό δυναμικό και το έργο που έχει επιτελέσει, με απώτερο σκοπό την ανάδειξη της επιστημονικής ευημερίας και ανάπτυξης. Χρησιμοποιεί την ανάλυση αναφορών και άλλες ποσοτικές μεθόδους, για να καταγράφει την πορεία μίας επιστημονικής έρευνας.

- **Cybermetrics**. Αφορά τη μελέτη των ποσοτικών πτυχών της κατασκευής και της χρήσης πληροφοριακών πόρων, δομών και τεχνολογιών, σε όλο το διαδίκτυο, με βάση τις βιβλιομετρικές και πληροφοριακές προσεγγίσεις. Το υποπεδίο αυτό, αφορά κυρίως την ανάλυση ιστοσελίδων, αντιμετωπίζοντάς τες ως έγγραφα. Όπως φαίνεται και στο Σχ. 3.1, υπερβαίνει ελάχιστα τα όρια του υποπεδίου "Bibliometrics", διότι ορισμένες δραστηριότητες δεν

καταγράφονται, αλλά επικοινωνούν συγχρονισμένα.

- **Webometrics**. Τέλος, το υποπεδίο αυτό, συμβάλλει με περαιτέρω μεθοδολογικές εξελίξεις και καλύπτεται πλήρως από το υποπεδίο "Bibliometrics", επειδή τα έγγραφα ιστού, είτε κείμενο είτε πολυμέσα, είναι καταγεγραμμένες πληροφορίες αποθηκευμένες σε διακομιστές. Επίσης, ολόκληρο το "Webometrics", περιλαμβάνεται στο "Cybermetrics" και συχνά οι ορολογίες των τελευταίων 2 υποπεδίων, χρησιμοποιούνται ως συνώνυμα.



Σχήμα 3.1: Σχέσεις μεταξύ των μετρικών επιστημών (Πηγή: [50])

3.3 Κατανόηση Μετρικών Έρευνας

Οι επιστημονικές μετρήσεις, αποτελούν τη λύση στην αναζήτηση της ποσοτικής μέτρησης του αντίκτυπου ενός άρθρου, ενός περιοδικού ή ενός συγγραφέα. Ειδικότερα, σε μία εποχή άφθονης πληροφόρησης και δημοσιευμένων βιβλιογραφιών, η ανάγκη για την καλύτερη κατανόηση του αντίκτυπου μίας επιστημονικής έρευνας, γίνεται ολοένα και μεγαλύτερη. Κάθε μετρική που χρησιμοποιείται στους τομείς που αναφέραμε και είδαμε στο Σχ. 3.1, προσφέρει μία διαφορετική οπτική γωνία στον τρόπο αξιολόγησης του αντίκτυπου . Βέβαια, πολλές φορές οι μετρικές έρευνας είναι αμφισβητήσιμες, καθώς προσπαθούν να καλύψουν πολυδιάστατες έννοιες. Η καλύτερη επιλογή είναι ο συνδυασμός της ποσοτικής και της ποιοτικής έρευνας, αλλά και η χρήση ποικίλων μετρικών, με σκοπό την απόκτηση μίας γενικής εικόνας που θα βοηθήσει στην τελική αξιολόγηση.

3.3.1 Σε Επίπεδο ενός Περιοδικού

Τα μετρικά ενός περιοδικού, χρησιμοποιούνται για να καθορίσουν τον αντίκτυπο που έχει στην επιστημονική κοινότητα. Παρακάτω, παρατίθενται κάποια από τα κυριότερα και ευρέως γνωστά μετρικά αυτής της κατηγορίας.

- **Impact Factor**. Ο παράγοντας αυτός υπολογίζεται για μία συγκεκριμένη χρονιά, ως το συνολικό άθροισμα των αναφορών που έλαβαν αυτή τη χρονιά, τα άρθρα που δημοσιεύθηκαν

στο συγκεκριμένο περιοδικό τα προηγούμενα 2 χρόνια, διαιρούμενο από το άθροισμα των άρθρων που δημοσιεύτηκαν εντός αυτών των 2 προηγούμενων ετών. Είναι η μετρική που χρησιμοποιείται πιο πολύ. Έχει δεχθεί πολλές αρνητικές κριτικές, καθώς στον αριθμητή υπολογίζονται οι αναφορές όλων των τύπων εγγράφων των περιοδικών, με εξαίρεση κάποια στοιχεία στον παρονομαστή, όπως οι επιστολές. Αυτό έχει ως αποτέλεσμα, περιοδικά με ενδιαφέρον τμήμα αλληλογραφίας να έχουν "φουσκωμένο" αυτόν τον παράγοντα αντίκτυπου , μέχρι και 75%. Αξίζει να σημειωθεί, ότι ορίζεται και ο παράγοντας για μία χρονιά, με βάση τα προηγούμενα 5 έτη.

- **Eigenfactor**. Στόχος είναι η μέτρηση της επίδρασης ενός περιοδικού, εντός ακαδημαϊκής κοινότητας. Μετράει το συνολικό αριθμό αναφορών που έλαβε ένα περιοδικό, εντός μίας πενταετίας. Αποτελεί βασικό μέτρο για το πόσοι άνθρωποι διαβάζουν ένα περιοδικό και για το πόσο θεωρούν ότι η περιοχή που καλύπτει, είναι σημαντική και ενδιαφέρουσα.

- **h5 Index**. Είναι ο μεγαλύτερος αριθμός h, έτσι ώστε τα άρθρα που δημοσιεύονται τα τελευταία 5 χρόνια, να έχουν τουλάχιστον h αναφορές το καθένα. Για παράδειγμα, αν ο δείκτης ισούται με 30, σημαίνει ότι το περιοδικό αυτό έχει δημοσιεύσει 30 άρθρα τα προηγούμενα 5 έτη που έχουν 30 ή περισσότερες αναφορές το καθένα.

3.3.2 Σε Επίπεδο ενός Άρθρου

Οι μετρικές που βασίζονται σ' ένα δημοσιευμένο άρθρο (αναφέρονται και ως "Altmetrics"), αποτελούν μία καινούρια προσέγγιση, για την ποσοτικοποίηση του αντίκτυπου μίας επιστημονικής δημοσιευμένης έρευνας. Προσπαθούν να υιοθετήσουν νέες πηγές δεδομένων, σε συνδυασμό με τα πιο παραδοσιακά μέτρα, όπως το "impact factor", για να μπορέσουν να κατανοήσουν τον τρόπο με τον οποίο ένα άρθρο χρησιμοποιείται, διαδίδεται και συζητείται. Με άλλα λόγια αποτελούν την εργαλειοθήκη ετερογενών σημείων δεδομένων, προσφέροντας την ικανότητα να κατανοούμε την επίδραση των δημοσιευμένων ερευνών, τόσο εντός όσο και εκτός της επιστημονικής κοινότητας. Οι παράγοντες της αμεσότητας και της κοινωνικοποίησης που συνοδεύουν τις συγκεκριμένες μετρικές, προσφέρουν έναν διαφορετικό τρόπο χαρτογράφησης της πορείας μίας νέας θεωρίας που έλαβε χώρα μέσα στην ακαδημαϊκή κοινότητα.

3.3.3 Σε Επίπεδο ενός Συγγραφέα

Στο συγκεκριμένο πεδίο, οι μετρικές αφορούν τον αντίκτυπο που έχει ένας συγκεκριμένος συγγραφέας, στην επιστημονική έρευνα. Είναι ένας νέας κλάδος της βιβλιομετρίας, του οποίου τα μετρικά χρησιμοποιούνται και από την επιστημονομετρία. Η πνευματική δραστηριότητα του συγγραφέα "ποσοτικοποιείται" με αυτούς τους μετρικούς δείκτες. Αυτές οι μετρήσεις ομαδοποιούνται σε 5 σύνολα: βιβλιομετρία (δημοσίευση και αναφορές), χρήση, συμμετοχή, αξιολόγηση, κοινωνική συνδεσιμότητα και σύνθετοι δείκτες. Έχει συντελέσει σε μεγάλο βαθμό στην επιθυμία των ερευνητών, τόσο για τη γνώση όσο και για την αναγνώριση.

- **h-index**. Ο δείκτης αυτός, αναφέρεται στον αριθμό των δημοσιεύσεων για τις οποίες ο συγγραφέας έλαβε αναφορές από άλλους συγγραφείς, τουλάχιστον τις ίδιες ή και παραπάνω

φορές από τον αριθμό αυτό. Εάν για παράδειγμα, ο δείκτης ισούται με 7, αυτό σημαίνει ότι ο συγγραφέας έχει δημοσιεύσει 7 άρθρα, όπου το καθένα απ' αυτά, έχει λάβει τουλάχιστον 7 αναφορές. Βέβαια, ο δείκτης δεν συνυπολογίζει τον αριθμό των συγγραφέων ανά άρθρο, οπότε άρθρα με πολλούς συγγραφείς, υπολογίζονται το ίδιο με τα άρθρα που έχουν έναν μόνο συγγραφέα.

- **g-index**. Βασίζεται στην κατανομή των αναφορών ενός συγγραφέα, έτσι ώστε με δεδομένο ένα σύνολο άρθρων που κατατάσσονται σε φθίνουσα σειρά, με βάση τον αριθμό αναφορών, ο δείκτης g, είναι ο μεγαλύτερος αριθμός, όπου τα g κορυφαία άρθρα έλαβαν αθροιστικά τουλάχιστον g^2 αναφορές. Η διαφορά μεταξύ των συγγραφέων γίνεται πλέον εμφανής. Αν και ο συγκεκριμένος δείκτης δεν χρησιμοποιείται τόσο συχνά όσο ο δείκτης h, θεωρείται όμως μία βελτίωση του τελευταίου, καθώς λαμβάνει υπ' όψιν τις αναφορές των κορυφαίων άρθρων.

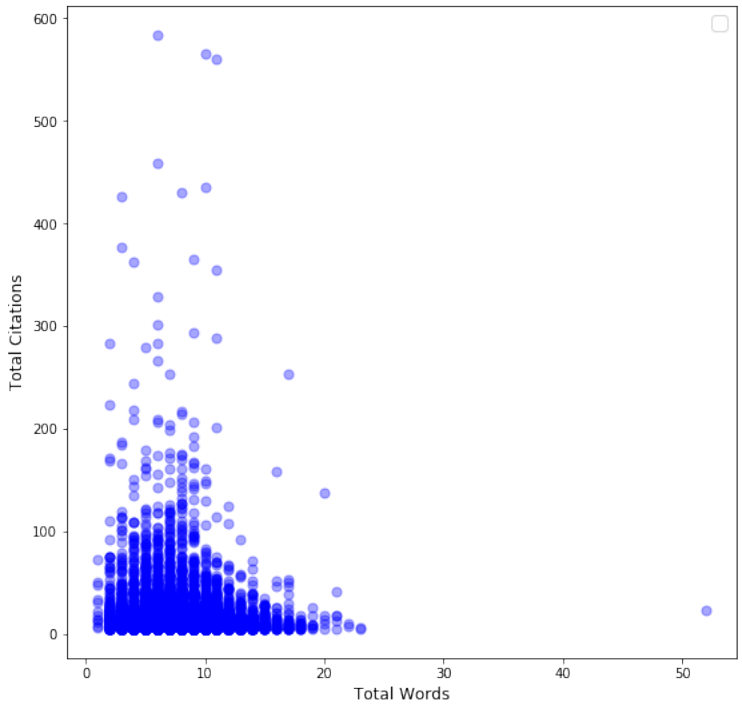
- **i10-index**. Αναφέρεται στον αριθμό των άρθρων που έχουν λάβει 10 ή και περισσότερες αναφορές και αποτελεί ένα απλό μετρικό. Άλλες μετρικές σε επίπεδο συγγραφέα, είναι ο συνολικός αριθμός δημοσιεύσεων του και ο συνολικός αριθμός αναφορών για όλες τις δημοσιεύσεις του.

3.4 Παράγοντες που Επηρεάζουν το Πλήθος των Αναφορών

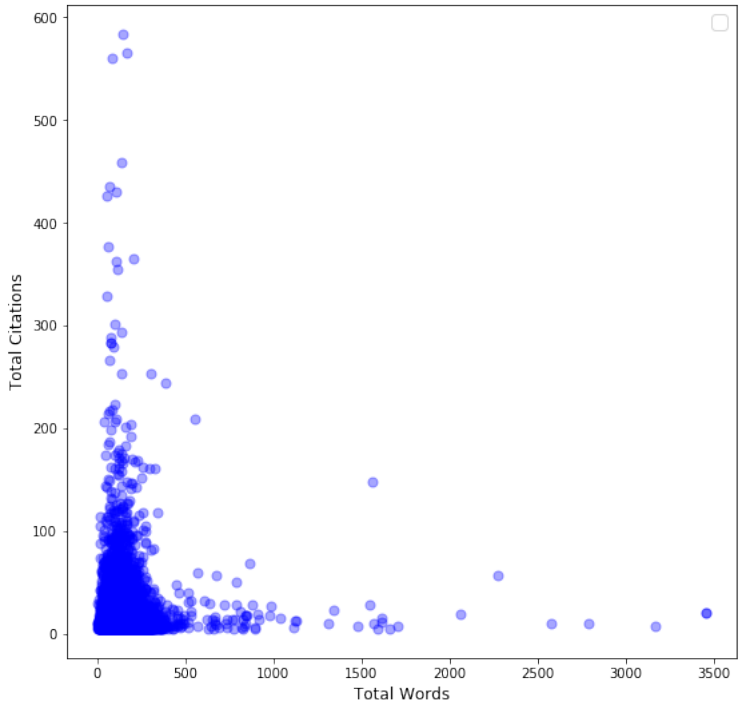
Γενικότερα, οι συγγραφείς αναζητούν τους καλύτερους δυνατούς τρόπους για να λάβουν μελλοντικά, όσο το δυνατόν περισσότερες αναφορές. Ένας αρκετά μεγάλος αριθμός μελετών, δείχνει ότι δεν είναι λίγοι οι παράγοντες που καθορίζουν τη μελλοντική αναγώριση ενός επιστημονικού άρθρου ή μίας επιστημονικής έρευνας. Υπάρχουν πολλές ακαδημαϊκές δημοσιεύσεις που λαμβάνουν λίγες αναφορές και υπάρχει ένα αρκετά μικρότερο ποσοστό δημοσιεύσεων που αναφέρονται σε μεγάλο βαθμό. Δεν υπάρχει ένα συγκεκριμένο "μονοπάτι" που θα πρέπει να ακολουθήσουν οι συγγραφείς, ούτως ώστε να αποκτήσουν τη μεγαλύτερη δυνατή αναγνώριση και επιτυχία. Οι παράγοντες είναι πολλοί και ο αυξανόμενος όγκος της δημοσιευμένης έρευνας, καθιστά το περιβάλλον ακόμα πιο δυσχερές. Με βάση μία έρευνα [52], όπου ανακτήθηκαν 2087 έγγραφα, μεταξύ των οποίων τα 198 ήταν μελέτες, αναγωρίστηκαν 28 παράγοντες που κατατάσσονται σε 3 βασικές κατηγορίες και επηρεάζουν τον αριθμό αναφορών. Παρακάτω, θα δούμε αυτές τις 3 κατηγορίες και κάποιους από τους βασικούς παράγοντες αυτών.

- **Έγγραφο**. Στην κατηγορία αυτή, κύριος συντελεστής στον αριθμό αναφορών, αποτελεί η καινοτομία και το ενδιαφέρον του θέματος της ίδιας της έρευνας. Τα πεδία και τα χαρακτηριστικά μελέτης, συγκαταλέγονται και αυτά στους παράγοντες επίδρασης στο πλήθος των αναφορών. Ακολουθεί ο σχεδιασμός της μελέτης και της μεθοδολογίας, καθώς και τα αποτελέσματα και ο τρόπος παρουσίασής τους, με τρόπο κατανοητό στον ακαδημαϊκό κόσμο. Ένας ακόμη παράγοντας, είναι η χρήση σχημάτων και προσαρτημάτων στο έγγραφο. Σημαντικό ρόλο επίσης, διαδραματίζει η "ηλικία" του εγγράφου και η έκτασή του σε μέγεθος. Τέλος, σε αυτή την κατηγορία, είναι και τα χαρακτηριστικά των τίτλων και των περιλήψεων που αποτελούν αναπόσπαστο κομμάτι του εγγράφου. Αξίζει να σημειωθεί, ότι μέρος

των επιστημονομετρικών δεδομένων της παρούσας εργασίας, αξιοποιήθηκαν για να εντοπίσουμε τη σχέση του μήκους των τίτλων και των περιλήψεων των εγγράφων, σχετικά με το συνολικό αριθμό αναφορών, όπως θα δούμε και στα επόμενα σχήματα. Στο Σχ. 3.2 και στο Σχ. 3.3, βλέπουμε το μέγεθος των τίτλων και των περιλήψεων αντίστοιχα, έναντι του πλήθους των αναφορών. Παρατηρούμε για τους τίτλους, ότι τα έγγραφα που έλαβαν για μία συγκεκριμένη χρονική περίοδο, υψηλό αριθμό αναφορών (300-600), είχαν μήκος τίτλου από 5 μέχρι 12 λέξεις. Αντίστοιχα, για τις περιλήψεις, τα έγγραφα που έλαβαν πολλές αναφορές την ίδια χρονική περίοδο, είχαν μέγεθος λέξεων από 100 έως 300 λέξεις. Βέβαια, υπάρχουν και έγγραφα, με μικρό μήκος τίτλου ή μικρό πλήθος λέξεων στις περιλήψεις τους που δεν έλαβαν πολλές αναφορές, γεγονός που μας δείχνει ότι βάση για την επιτυχία των εγγράφων, δεν αποτελούν μόνο τα καθαυτού χαρακτηριστικά του, αλλά και άλλοι παράγοντες.



Σχήμα 3.2: Σύνολο αναφορών έναντι συνολικού πλήθους λέξεων των τίτλων



Σχήμα 3.3: Σύνολο αναφορών έναντι συνολικού πλήθους λέξεων των περιλήψεων

- **Περιοδικό**. Κύριος παράγοντας αυτής της κατηγορίας, είναι το "impact factor" του περιοδικού που αναφέραμε και στην Υποενότητα 3.3.1. Άλλοι παράγοντες της κατηγορίας που διαδραματίζουν σημαντικό ρόλο στο πλήθος των αναφορών, είναι η γλώσσα του περιοδικού καθώς και τα πεδία μελέτης και εφαρμογής του. Τέλος, το κύρος και η φήμη του περιοδικού (διαφορετικό από το "impact factor"), η δημοσίευση παραπλήσιων άρθρων, καθώς και η διαφήμιση που επιλέγεται για την προώθηση των εγγράφων, συγκαταλέγονται στους παράγοντες επιρροής.

- **Συγγραφέας**. Η τρίτη και τελευταία κατηγορία, αφορά τον συγγραφέα και τη φήμη που έχει αποκτήσει στον ακαδημαϊκό χώρο. Ο αριθμός των δημιουργών μίας δημοσιευμένης έρευνας, καθώς και η επιστημονική κατάταξη αυτών, καθιστά τα έγγραφα άμεσα συνδεδεμένα με αυτούς τους συντελεστές. Επιπρόσθετα, το φύλο, η ηλικία και η παραγωγικότητα του συγγραφέα, φαίνεται να αποτελούν αναπόσπαστο κομμάτι των συντελεστών που σχετίζονται με το πλήθος των παραπομπών. Φυσικά, δεν θα μπορούσε να απουσιάζει ο συντελεστής της χρηματοδότησης και οι αυτο-αναφορές (παραπομπές των άρθρων από το ίδιο το περιοδικό), των ίδιων των δημιουργών.

Παρ 'όλα αυτά, η μελετημένη στατηγική και παρουσίαση των εγγράφων, είναι συμπληρωματική και δεν υποκαθιστά την ανάγκη για δημιουργία ερευνών υψηλής ποιότητας. Συγγραφείς που επιδιώκουν να είναι ευρέως αναγνωρίσιμοι και να λαμβάνουν μελλοντικές αναφορές, θα πρέπει να γνωρίζουν ότι το πλήθος αναφορών είναι μόνο μία πτυχή της έρευνας. Πρέπει να επιδιώκουν να γράφουν πλήρη και ουσιαστικά άρθρα, με κύριο προτέρημα την αμεροληψία και την ορθότητα

τους. Συνυπολογίζοντας τα παραπάνω, σε συνδυασμό με τους προαναφερθέντες παράγοντες, όπως ο υψηλός αντίκτυπος ενός περιοδικού, μπορούν να προβαίνουν στη συγγραφή και στη δημοσίευση των ερευνών τους.

Κεφάλαιο 4

Επιστημονομετρία και Μηχανική Μάθηση

Αυτό το κεφάλαιο περιλαμβάνει κυρίως τη θεωρία που είναι άμεσα συνδεδεμένη με τη διεξαγωγή των πειραμάτων. Πριν από κάθε πείραμα, αποσαφηνίζεται το θεωρητικό υπόβαθρο πίσω από αυτό. Αναλύονται τα μοντέλα που κατασκευάστηκαν και αφορούν τη μοντελοποίηση θεμάτων, τη συσταδοποίηση εγγράφων, αλλά και την πρόβλεψη μελλοντικών αναφορών που θα λάβει το κάθε έγγραφο την επόμενη τριετία. Επίσης, παρατίθενται κάποιες τεχνικές που αφορούν τη γενικότερη βελτιστοποίηση κάποιων εκ των αλγορίθμων που εφαρμόστηκαν.

4.1 Παρουσίαση των Επιστημονομετρικών Δεδομένων

4.1.1 Εισαγωγή

Ένα σύνολο δεδομένων, είναι μία συλλογή δεδομένων που αντιστοιχεί στο περιεχόμενο ενός πίνακα βάσης δεδομένων ή ενός ενιαίου πίνακα στατιστικών δεδομένων, όπου η κάθε στήλη του πίνακα αντιπροσωπεύει ένα συγκεκριμένο γνώρισμα (ή αλλιώς μεταβλητή) και η κάθε σειρά του (ή αλλιώς εμφάνιση), αντιστοιχεί σ' ένα μέλος του εν λόγω συνόλου δεδομένων.

Ένα από τα δυσκολότερα και πιο χρονοβόρα προβλήματα στη Μ.Μ για τη δημιουργία αποτελεσματικών μοντέλων, είναι η απόκτηση των σωστών δεδομένων, στη σωστή μορφή. Τα δεδομένα που θα χρειαστεί να αποκτήσει ο κάθε χρήστης, θα πρέπει να σχετίζονται με τα αποτελέσματα που θέλει να προβλέψει, ενώ η ποικιλία των προβλημάτων της Μ.Μ, καθιστά ακόμα πιο δύσκολη τη διαδικασία της απόκτησης ενός αντιπροσωπευτικού συνόλου δεδομένων. Αποτελεί την πιο κρίσιμη πτυχή που καθιστά εφικτή την αποτελεσματική εκπαίδευση των αλγορίθμων. Τα δεδομένα που θα αποτελέσουν τη βάση των μοντέλων της Μ.Μ, πρέπει να διέπονται από αντικειμενικότητα και να μην χαρακτηρίζονται από απώλειες και ανωμαλίες.

4.1.2 Τα Δεδομένα

Τα επιστημονομετρικά δεδομένα [13] που χρησιμοποιήθηκαν για τη διεξαγωγή πειραμάτων, συλλέχθηκαν το Μάϊο του 2010, διαθέτουν 7 γνωρίσματα και αποτελούνται από 629.813 εμφανί-

σεις. Υπάρχουν πάνω από 632.752 σχέσεις αναφορών μεταξύ των εγγράφων, οι οποίες χρειάστηκε να εντοπιστούν, ώστε να μπορέσουμε να έχουμε και το συνολικό αριθμό αναφορών που έλαβε ένα έγγραφο την κάθε χρονιά, από τη χρονιά δημοσίευσής του μέχρι το 2009, καθώς ήταν απαραίτητος για τα πειράματα. Η χρονιά 2010 δεν λήφθηκε υπ' όψιν, γιατί τότε τα περισσότερα έγγραφα είχαν μηδενικό σύνολο αναφορών. Η παλαιότερη χρονολογία δημοσίευσης ενός εγγράφου, εντοπίζεται το 1900, ενώ το μεγαλύτερο πλήθος αναφορών που συγκέντρωσε ένα έγγραφο, είναι 816. Στον Πίνακα 4.1 παρατίθεται η επεξήγηση των συμβόλων που υπήρχαν στο σύνολο γνωρισμάτων των επιστημονομετρικών δεδομένων και εν συνεχεία, δίνεται ένα παράδειγμα για την καλύτερη κατανόησή τους.

Σύμβολα	Επεξήγηση
♯*	Τίτλος εγγράφου
♯@	Συγγραφείς
♯t	Χρονιά έκδοσης
♯c	Περιοδικό έκδοσης
♯index	Αναγνωστικό (id) αυτού του εγγράφου
♯%	Τα αναγνωριστικά των εγγράφων που έκανε αναφορά το συγκεκριμένο έγγραφο. Μπορεί να αποτελείται από πολλές γραμμές, όπου η κάθε μία αντιπροσωπεύει και το αναγννωριστικό του εγγράφου.
♯!	Περίληψη

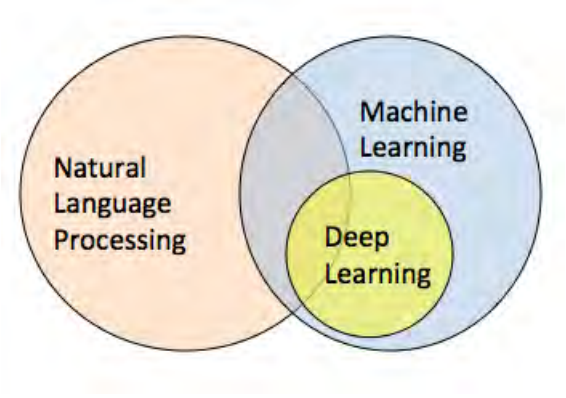
Πίνακας 4.1: Παράθεση και επεξήγηση γνωρισμάτων των επιστημονομετρικών δεδομένων

- ♯*Information geometry of U-Boost and Bregman divergence
- ♯@Noboru Murata,Takashi Takenouchi,Takafumi Kanamori,Shinto Eguchi
- ♯t2004
- ♯cNeural Computation
- ♯index436405
- ♯%94584
- ♯%282290
- ♯%605546
- ♯!We aim at an extension of AdaBoost to U-Boost, in the paradigm to build a stronger classification machine from a set of weak learning machines. A geometric understanding of the Bregman divergence defined by a generic convex function U leads to the U-Boost method in the framework of information geometry extended to the space of the finite measures over a label set.

4.2 Μοντελοποίηση Θεμάτων

4.2.1 Επεξεργασία Φυσικής Γλώσσας

Ο τομέας της Επεξεργασίας Φυσικής Γλώσσας (Ε.Φ.Γ), είναι ένα πεδίο που συνδυάζει την Τ.Ν και τη γλωσσολογία, προσφέροντας τη δυνατότητα στους υπολογιστές να είναι ικανοί να μαθαίνουν και να κατανοούν την ανθρώπινη γλώσσα. Έχει ευρείες εφαρμογές στο χώρο της επιστήμης, όπως είναι η ανάλυση συναισθημάτων, η μοντελοποίηση θεμάτων, η ανίχνευση ανεπιθύμητων μηνυμάτων, η αναγνώριση ομιλίας κ.λπ. Οι περισσότερες τεχνικές που χρησιμοποιούνται σ' αυτόν τον τομέα, βασίζονται στη Μ.Μ, με απώτερο σκοπό την εξαγωγή νοήματος από την ανθρώπινη γλώσσα. Αυτός είναι και ο λόγος που υπάρχει αλληλεπικάλυψη της Μ.Μ με την Ε.Φ.Γ, όπως φαίνεται και στο Σχ. 4.1. Στο ακόλουθο σχήμα, δεν υπάρχει η συνιστώσα της γλωσσολογίας που αναφέραμε προηγουμένως, παρόλο που αποτελεί ισχυρό παράγοντα και απαιτεί τη βαθιά κατανόηση του τρόπου με τον οποίο χρησιμοποιούμε τη γλώσσα. Ο τρόπος γραφής σε διαφορετικά είδη, όπως για παράδειγμα τα ερευνητικά άρθρα ή οι επιστολές, καθιστά την ανάγκη εμπέδωσης των παραγόντων που διέπουν τη συνιστώσα της γλωσσολογίας, έτσι ώστε να μπορούμε να κωδικοποιήσουμε και να εφαρμόσουμε το πρόβλημα μας, σ' έναν αλγόριθμο Μ.Μ. Ο ρόλος της Μ.Μ και της Τ.Ν στο πεδίο της Ε.Φ.Γ, είναι να βελτιώνει, να επιταχύνει και να αυτοματοποιεί τις λειτουργίες που αφορούν τις τεχνικές ανάλυσης κειμένων.



Σχήμα 4.1: Ε.Φ.Γ, Μ.Μ και Β.Μ (Πηγή: [42])

4.2.2 Ταξινόμηση Κειμένου

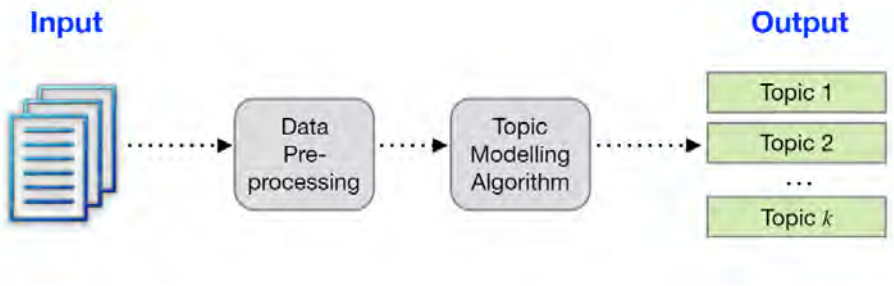
Η ταξινόμηση κειμένου, είναι μία διαδικασία μέσω της οποίας εκχωρούνται ετικέτες ή κατηγορίες στα κείμενα, ανάλογα με τα περιεχόμενά τους. Η ταξινόμηση κειμένου ονομάζεται επίσης κατηγοριοποίηση κειμένου ή ταξινόμηση εγγράφων ή αλλιώς κατηγοριοποίηση εγγράφων. Τέτοια κείμενα μπορεί να είναι σχόλια στα κοινωνικά μέσα, κριτικές, πολιτικές ομιλίες, ακόμη και νομικά ή ιατρικά έγγραφα. Με τους ταξινομητές κειμένου, δίνεται η δυνατότητα για την οργάνωση, τη δομή και την κατηγοριοποίηση σχεδόν του οτιδήποτε. Αυτοί οι ταξινομητές, λαμβάνουν ως είσοδο κείμενο, επεξεργάζονται και αναλύουν τα περιεχόμενα του κειμένου και στη συνέχεια είναι ικανοί να παράγουν αυτόματα, σχετικές με το θέμα, ετικέτες.

Υπάρχουν 2 προσεγγίσεις ταξινόμησης κειμένου, η μία είναι η χειροκίνητη ταξινόμηση και η άλλη είναι η αυτόματη. Στην πρώτη προσέγγιση, πραγματοποιείται από τον ίδιο τον σχολιαστή η κατηγοριοποίηση και μπορεί να προσφέρει ποιοτικά αποτελέσματα, αλλά είναι χρονοβόρα και δαπανηρή. Η δεύτερη, με την οποία ασχοληθήκαμε και στην παρούσα διπλωματική, αφορά τη Μ.Μ, την Ε.Φ.Γ και άλλες μεθόδους για τον αυτόματο εντοπισμό κρυφών δομών στα δεδομένα. Είναι ταχύτερη από την πρώτη και πιο οικονομική.

Υπάρχουν αρκετές μέθοδοι στην τελευταία προσέγγιση, οι οποίες μπορούν να κατηγοριοποιηθούν σε 3 βασικούς τύπους συστημάτων. Στα συστήματα βασισμένα σε κανόνες, σ' αυτά που στηρίζονται στη Μ.Μ και στα υβριδικά συστήματα. Στον πρώτο τύπο, το κείμενο ταξινομείται σε οργανωμένες ομάδες, βασισμένες σ' ένα σύνολο γλωσσολογικών κανόνων. Η δημιουργία κατηγοριών, πραγματοποιείται με τον εντοπισμό σχετικά όμοιων στοιχείων ενός κειμένου, με βάση το περιεχόμενο που το χαρακτηρίζει. Στον δεύτερο τύπο, οι ταξινομήσεις βασίζονται σε προηγούμενες παρατηρήσεις. Ένας αλγόριθμος της Μ.Μ, προπονείται με δεδομένα εκπαίδευσης που είναι σε ζεύγη συνόλων χαρακτηριστικών και γνωρίζοντας μία συγκεκριμένη έξοδο (ετικέτες), την αντιστοιχεί σε μία συγκεκριμένη είσοδο. Στον συγκεκριμένο τύπο, η εξαγωγή χαρακτηριστικών που προηγείται της εκπαίδευσης του αλγορίθμου, μετατρέπει το κείμενο σε αριθμητική αναπαράσταση, με τη μορφή ενός διανύσματος. Η πιο ευρέως γνωστή προσέγγιση για την εξαγωγή χαρακτηριστικών, είναι το μοντέλο Bag of Words, το οποίο θα αναλύσουμε και στην επόμενη υποενότητα. Γι' αυτόν τον λόγο, τα δεδομένα εκπαίδευσης είναι σε ζεύγη, όπως αναφέραμε προηγουμένως και τροφοδοτούνται στο μοντέλο με απώτερο σκοπό την ταξινόμηση κειμένου, ακόμα και σε δεδομένα που δεν έχει "ξανασυναντήσει". Τέλος, ο τύπος που βασίζεται στα υβριδικά συστήματα, συνδυάζει τους δύο προηγούμενους τύπους, με σκοπό την περαιτέρω βελτίωση της αποδοτικότητας του μοντέλου. Η προσθήκη ειδικών κανόνων, καθιστά τον τελευταίο τύπο, εύκολα ρυθμιζόμενο, για ετικέτες που δεν έχουν καθοριστεί σωστά από τον ταξινομητή βάσης.

Η ταχεία ανάπτυξη των δεδομένων κειμένου, συχνά υπερβαίνει το όριο του τι μπορεί ένα άτομο να διαβάσει ή να επεξεργαστεί. Θα προσπαθήσουμε να κατανοήσουμε από τις περιλήψεις των εγγράφων των επιστημονομετρικών μας δεδομένων, τις γενικές τάσεις που περιέχουν αυτά και να εξάγουμε σαφή, ουσιαστικά διαχωρισμένα και καλής ποιότητας θέματα. Αυτή η διαδικασία που απεικονίζεται και στο Σχ. 4.2, καλείται μοντελοποίηση θεμάτων και είναι ένα συνηθισμένο παράδειγμα της ταξινόμησης κειμένου. Ο σκοπός της είναι η αυτόματη ανακάλυψη μίας "κρυμμένης" θεματικής δομής, σ' έναν μεγάλο όγκο εγγράφων κειμένου. Η μοντελοποίηση θεμάτων και γενικότερα η ταξινόμηση κειμένου, ανήκει στην κατηγορία της μη εποπτευόμενης μάθησης. Επίσης, βοηθάει την ταξινόμηση κειμένου, τοποθετώντας παρόμοιες λέξεις μαζί στα θέματα, αντί να χρησιμοποιεί την κάθε λέξη-όρο ως χαρακτηριστικό γνώρισμα.

Συνοψίζοντας, τα μοντέλα θεμάτων βοηθούν στην ανάλυση των κειμένων για τον εντοπισμό "κρυμμένων" θεμάτων που βασίζονται σε λέξεις, οι οποίες εμφανίζονται μαζί ή ταυτόχρονα μέσα στο κείμενο. Γενικότερα, οι αναπτύξεις μοντέλων στον πραγματικό κόσμο, απαιτούν εντατική επαλήθευση των ειδικών και μία διαρκής προσπάθεια για τη βελτίωση τους. Ειδικότερα, στον κόσμο της βιβλιογραφίας, συναντάται μία έλλειψη σχετικά με τις αναλύσεις του πραγματικού κόσμου, με χαρακτηριστικό παράδειγμα μίας τέτοιας ανάλυσης, η μοντελοποίηση θεμάτων.



Σχήμα 4.2: Διαδικασία μοντελοποίησης θεμάτων (Πηγή: [43])

4.2.3 Ο Αλγόριθμος LDA

Η μοντελοποίηση θεμάτων μπορεί να αποδειχθεί αρκετά ισχυρή σε πεδία όπως της ταξινόμησης κειμένου, του εντοπισμού θεμάτων σε κείμενα, των συστημένων συστημάτων που προτείνουν στον χρήστη για παράδειγμα, την ανάγνωση κάποιων άρθρων που είναι παρόμοια σε θέμα και δομή με άλλα που έχει ήδη διαβάσει, χρησιμοποιώντας μέτρα ομοιότητας κ.λπ. Όλα τα μοντέλα θεμάτων, βασίζονται στην ίδια υπόθεση, ότι κάθε έγγραφο αποτελείται από ένα "μίγμα" άλλων θεμάτων και κάθε θέμα αποτελείται από μία συλλογή λέξεων. Υπάρχουν διάφοροι αλγόριθμοι που αφορούν τη μοντελοποίηση θεμάτων και οι πιο ευρέως γνωστοί παρατίθενται παρακάτω.

- **Latent Dirichlet Allocation (LDA)**. Είναι το μοντέλο που χρησιμοποιήθηκε για τη μοντελοποίηση θεμάτων των επιστημονομετρικών μας δεδομένων. Είναι ένα γενετικό πιθανοτικό μοντέλο για σύνολα διακριτών δεδομένων, όπως είναι τα κείμενα.

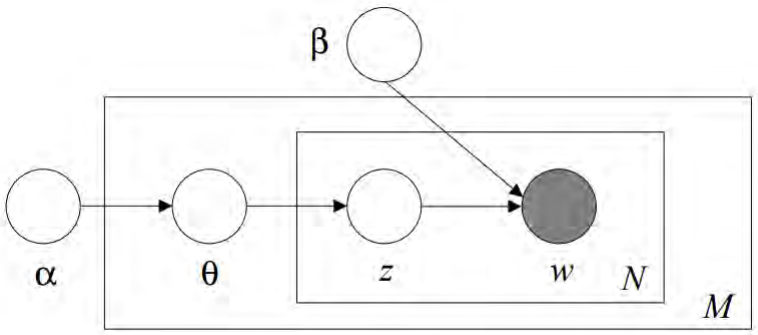
- **Latent Semantic Analysis ή Latent Semantic Indexing (LSI)**. Ο αλγόριθμος LSI, είναι ένας μαθηματικός αλγόριθμος που χρησιμοποιείται για να εντοπίσει το επίπεδο σημαντικότητας και σχετικότητας ενός τμήματος κειμένου ή να καθορίσει κατά πόσο σημαντικό ή σχετικό είναι αυτό το κείμενο με το θέμα, την επικεφαλίδα ή τις λέξεις (κλειδιά) της αναζήτησης. Χρησιμοποιείται κυρίως από τις μηχανές αναζήτησης του διαδικτύου και είναι αρκετά πολύπλοκος.

- **Probabilistic Latent Semantic Indexing (pLSI)**. Αποτελεί ένα πιο εκφραστικό μοντέλο από το LSI, επιτρέποντας διάφορα θέματα ανά έγγραφο, σε διάφορες αναλογίες. Επίσης, αποτελείται από πολλές παραμέτρους, γεγονός που οδηγεί κάποιες φορές, στο φαινόμενο της υπερ-προσαρμογής του εκάστοτε μοντέλου.

Ο αλγόριθμος LDA, είναι μία προσέγγιση που στηρίζεται στις κατανομές πολυωνυμικών πιθανοτήτων πάνω σε όρους που δημιουργούνται από τη συγκέντρωση όμοιων ομάδων λέξεων, οι οποίες βασίζονται στην ταυτόχρονη εμφάνιση μέσα σε έγγραφα, με παρόμοια θέματα. Σκοπός είναι η παραγωγή λογικών θεμάτων που πρέπει να επαληθευτούν από κάποιον ειδικό στον τομέα, για να διαπιστωθεί η αξιοπιστία τους, καθώς ελοχεύει ο κίνδυνος δημιουργίας αναξιόπιστων θεμάτων. Είναι ένα ιεραρχικό μοντέλο Bayesian, αποτελούμενο από 3 επίπεδα, όπου κάθε στοιχείο μίας συλλογής, διαμορφώνεται ως ένα πεπερασμένο "μίγμα", σ' ένα υποκείμενο σύνολο θεμάτων. Στη

συνέχεια, κάθε θέμα διαμορφώνεται ως ένα άπειρο "μίγμα" πάνω σ' ένα υποκείμενο σύνολο πιθανών θεμάτων. Στα πλαίσια μοντελοποίησης κειμένων, οι πιθανότητες του θέματος παρέχουν μία ρητή αναπαράσταση ενός εγγράφου. Με άλλα λόγια, τα έγγραφα είναι η κατανομή πιθανότητας πάνω σε "κρυμμένα" θέματα και τα θέματα είναι η κατανομή πιθανότητας πάνω στις λέξεις.

Ο χαρακτηρισμός "κρυμμένα", στηρίζεται στο γεγονός ότι τα θέματα γίνονται ορατά μόνο κατά τη διάρκεια της διαδικασίας της μοντελοποίησης, ενώ προηγουμένως είχαμε απλά ένα σύνολο λέξεων και εγγράφων. Η μέθοδος LDA, υποθέτει ότι υπάρχει ένα σταθερό σύνολο θεμάτων που το καθένα αποτελείται από ένα σύνολο λέξεων. Κάθε έγγραφο στο σύνολο των δεδομένων, μπορεί να περιγραφεί από ένα σύνολο θεμάτων και το κάθε θέμα από ένα σύνολο λέξεων. Οι λέξεις συνδέονται στα θέματα που αντιστοιχούν καλύτερα σ΄ αυτές και έπειτα, τα θέματα συνδέονται με τα έγγραφα βάσει των θεμάτων που το κάθε έγγραφο πραγματεύεται. Η λύση της μεθόδου, για παράδειγμα σε θέμα σχετικά με τον καιρό, είναι στην ακόλουθη μορφή: (0.4*υγρασία, 0,35*ηλιος, 0.2*σύννεφα). Αυτός είναι και ο λόγος που το κάθε θέμα είναι μία κατανομή λέξεων. Επίσης, η σειρά με την οποία το σύνολο των λέξεων εμφανίζεται μέσα στο έγγραφο, δεν επηρεάζει το μοντέλο LDA, ώστε να μπορεί ακόμα και με τις ίδιες λέξεις και με διαφορετική σειρά εμφάνισης, να "μαντέψει" το ίδιο θέμα. Για τη βαθύτερη κατανόηση του μοντέλου, θα συζητηθεί παρακάτω το μαθηματικό υπόβαθρο πίσω από τον αλγόριθμο LDA. Στο Σχ. 4.3, απεικονίζεται το γραφικό μοντέλο του LDA.



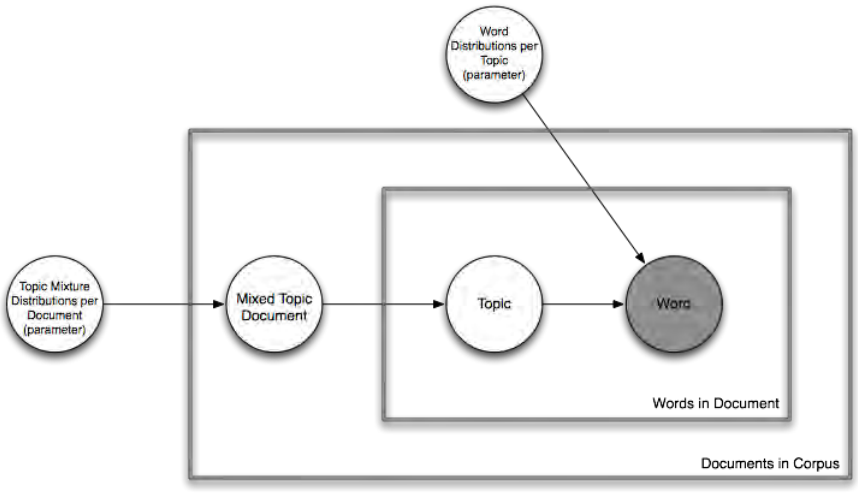
Σχήμα 4.3: Γραφικό μοντέλο του LDA - Σχέσεις μεταξύ των παραμέτρων του (Πηγή: [10])

,

Τα δύο σχηματιζόμενα ορθογώνια, *N* και *M*, συμβολίζουν το επίπεδο που πραγματοποιείται μία ενέργεια, δηλαδή αν είναι στο επίπεδο με τις λέξεις (*N*), στο επίπεδο με τα έγγραφα (*M*) ή και στα δύο. Ακολουθεί μία σύντομη επεξήγηση όλων των συμβόλων του αλγορίθμου LDA, κάποια από τα οποία απεικονίζονται στο παραπάνω σχήμα. Το α του Σχήματος 4.3, είναι μία σταθερά. Ουσιαστικά αντιπροσωπεύει πίνακα διαστάσεων *M* x *k*. Η σταθερά αυτή, διαμορφώνεται με την αναπαραγωγή της μοναδικής τιμής στον πίνακα, σε κάθε μεμονωμένο κελί. Στο Σχ. 4.4, τα σύμβολα αντικαθίστανται από λέξεις, για την καλύτερη κατανόηση του μοντέλου.

- **k**. Συμβολίζει τον αριθμό των θεμάτων που ένα έγγραφο ανήκει. Όπως αναφέραμε προηγουμένως, ο αριθμός θεμάτων είναι σταθερός.

- **V**. Μέγεθος λεξιλογίου
- **M**. Αριθμός εγγράφων
- **N**. Αριθμός των λέξεων που περιλαμβάνει το κάθε έγγραφο.
- **w**. Μία λέξη σ' ένα συγκεκριμένο έγγραφο. Εκπροσωπείται από ένα διάνυσμα μεγέθους *V*. Αποτελεί πίνακα διαστάσεων *V* x *N*.
- **D**. Συμβολίζει ολόκληρη τη συλλογή των *M* εγγράφων.
- **z**. Εκπροσωπεί κάθε θέμα που έχει αντιστοιχηθεί σε κάθε λέξη, δημιουργώντας κάθε έγγραφο, το οποίο είναι ένα σύνολο από *k* θέματα. Κάθε θέμα είναι μία κατανομή από λέξεις. Είναι πίνακας με διαστάσεις *k* x *N*.
- **α**. Είναι μία μοναδική παράμετρος-διάνυσμα μίας Dirichlet κατανομής, για κάθε έγγραφο και σχετίζεται με την κατανομή που διέπει την εμφάνιση της διανομής των θεμάτων, για κάθε έγγραφο της συλλογής εγγράφων.
- **θ**. Είναι ένας τυχαίος πίνακας θ (*i*,*j*) που ορίζει την πιθανότητα του *i*-οστού εγγράφου, να περιέχει λέξεις που ανήκουν στο *j*-οστό θέμα. Έχει διαστάσεις *k* x *M*.
- **β**. Το β έχει κάποια κατανομή, την λεγόμενη Dirichlet κατανομή. Με βάση αυτή την κατανομή, το β δημιουργεί *k* ξεχωριστές λέξεις για κάθε θέμα. Είναι ένας τυχαίος πίνακας β (*i*,*j*) που ορίζει την πιθανότητα του *i*-στού θέματος, να περιέχει την *j*-οστή λέξη. Οι διαστάσεις του συγκεκριμένου πίνακα είναι *V* x *k*.



Σχήμα 4.4: Επεξήγηση του γραφικού μοντέλου LDA (Πηγή: [55])

,

Το μοντέλο θεωρεί ότι κάθε έγγραφο είναι ένα "μίγμα" θεμάτων, όπου κατανέμονται οι διαστάσεις του "μίγματος" συνεχούς αποτίμησης, ως μία λανθάνουσα τυχαία μεταβλητή Dirichlet. Υποθέτει ότι τα νέα έγγραφα δημιουργούνται ορίζοντας αρχικά τον αριθμό των λέξεων σε κάθε

έγγραφο. Επιλέγει το "μίγμα" θεμάτων για το κάθε έγγραφο μέσω ενός σταθερού συνόλου θεμάτων, για παράδειγμα 20% θέμα *A*, 30% θέμα *B* και 50% για το θέμα *C*. Στη συνέχεια, παράγει λέξεις στο έγγραφο, επιλέγοντας αρχικά θέμα με βάση την κατανομή εγγράφου και μετά επιλέγει λέξη με βάση την κατανομή θέματος.

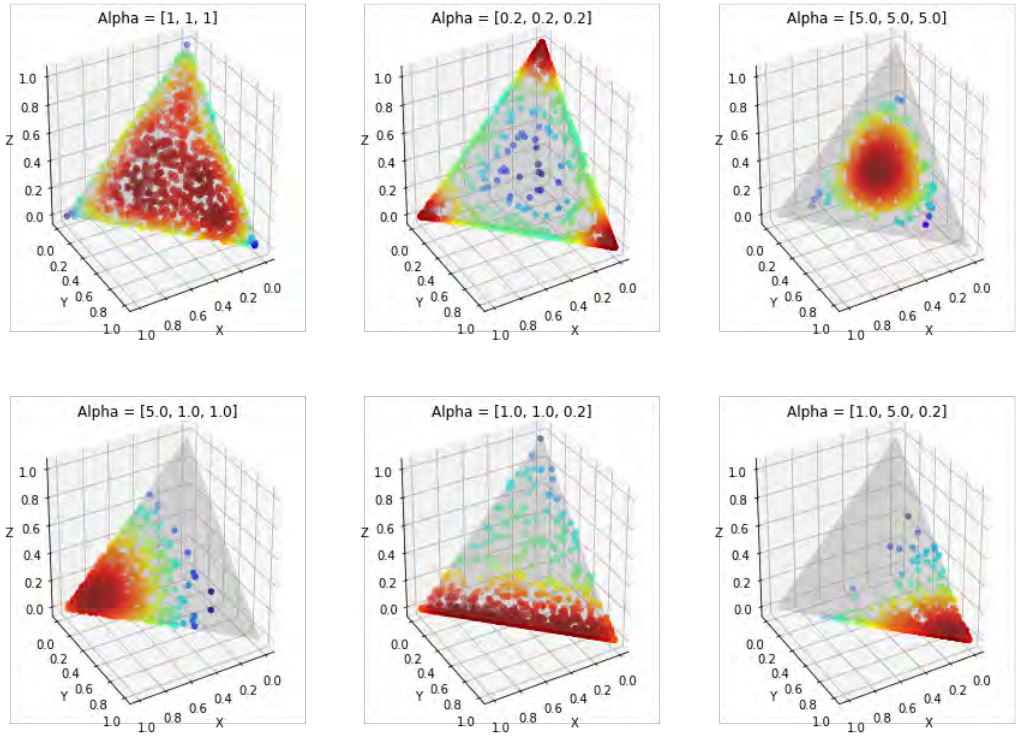
Αξίζει να αναφέρουμε στο σημείο αυτό, κάποια προτερήματα και ελαττώματα σχετικά με τον αλγόριθμο LDA. Αρχικά, είναι αποτελεσματικός στη μοντελοποίηση θεμάτων και εύκολος με την απόκτηση της αντίληψης σχετικά με τη λειτουργία του. Επιπρόσθετα, έχει επιδείξει αρκετά καλά αποτελέσματα σε διάφορους τομείς και χρησιμοποιείται σε διάφορες εφαρμογές. Είναι γρήγορος και μπορεί να προβλέπει θέματα για έγγραφα που δεν έχει "ξανασυναντήσει". Όσον αφορά τα ελαττώματα, θεωρείται ότι επιδέχεται πολλές ρυθμίσεις στις παραμέτρους του που συντελούν στην αποτελεσματικότητά του. Επίσης, πρέπει να είναι προκαθορισμένος ο αριθμός των παραγόμενων θεμάτων, αφού ο αλγόριθμος δεν μπορεί από μόνος του να γνωρίζει, τον ιδανικό αριθμό θεμάτων. Αυτός ήταν και ο λόγος για τις δοκιμές που πραγματοποιήθηκαν στο δικό μας μοντέλο [5.1.7], προκειμένου να διαπιστωθεί ένας αριθμός θεμάτων, όπου τα θέματα θα ήταν σαφή και αντιπροσωπευτικά. Επίσης, τα θέματα που εξάγονται, πρέπει πάντα να ερμηνεύονται από κάποιο ειδικό άτομο, με απώτερο σκοπό την καλύτερη κατανόηση αυτών. Σε περίπτωση που γνωρίζουμε ότι στα έγγραφά μας μπορεί να υπάρχει ένα θέμα το οποίο δεν εντόπισε ο αλγόριθμος, δεν υπάρχει τρόπος υπόδειξης του αλγορίθμου για το ποιες λέξεις θα πρέπει να είναι μαζί. Τέλος, η Dirichlet κατανομή, δεν είναι ικανή να κατανοήσει τις σχέσεις μεταξύ των θεμάτων που εντοπίζονται.

Ένα παράδειγμα μίας Dirichlet κατανομής, απεικονίζεται στο Σχ. 4.5. Συγκεκριμένα, είναι ένα πρόβλημα 3 διαστάσεων που επηρεάζει το σχήμα του θ. Για πρόβλημα με διαστάσεις N, έχουμε ένα διάνυσμα N μήκους. Όπως φαίνεται, η κατανομή του θ αλλάζει, τροποποιώντας τις τιμές του α. Για μικρές τιμές του α, η κατανομή ωθείται προς τις γωνίες του τριγώνου, ενώ για μεγάλες τιμές του α, η κατανομή "μαζεύεται" προς τη μέση. Οι κουκίδες των ίδιων χρωμάτων, συμβολίζουν τα έγγραφα που ανήκουν στο ίδιο θέμα.

4.3 Συσταδοποίηση Εγγράφων

4.3.1 Εισαγωγή

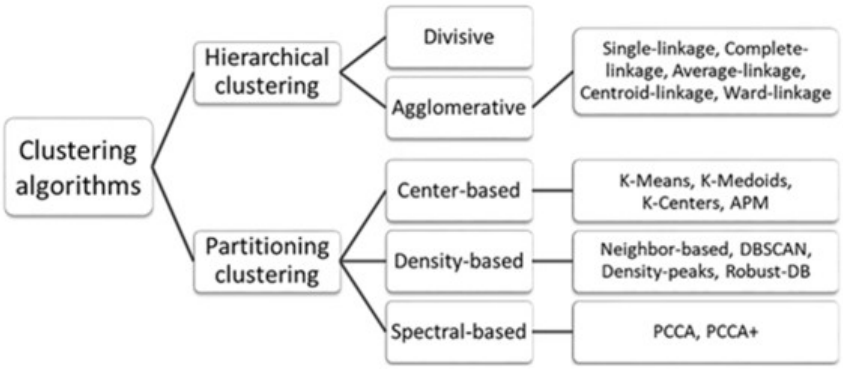
Η συσταδοποίηση ή αλλιώς ομαδοποίηση, είναι από τις πιο γνωστές μεθόδους της μη επoπτευόμενης μάθησης. Είναι μία τεχνική για την ανάλυση στατιστικών δεδομένων που χρησιμοποιείται σε πολλούς τομείς. Στην επιστήμη δεδομένων, η τεχνική της συσταδοποίησης χρησιμοποιείται για την απόκτηση χρήσιμων πληροφοριών από υπάρχοντα δεδομένα, παρατηρώντας ποια σημεία ανήκουν ή όχι, στις ίδιες ομάδες. Συγκεκριμένα, ασχολείται με την εύρεση δομής σε μία συλλογή μη επισημασμένων δεδομένων (απουσία ετικετών). Η κάθε συστάδα που δημιουργεί η συσταδοποίηση, είναι μία συλλογή απο αντικείμενα, τα οποία είναι όμοια μεταξύ τους και ανόμοια με τα αντικείμενα των άλλων σχηματιζόμενων συστάδων. Είναι μία διαδικασία πολύ χρήσιμη στον τομέα του κειμένου. Η συσταδοποίηση εγγράφων, έχει χρησιμοποιηθεί εκτενώς σε διάφορους τομείς εξόρυξης κειμένου και ανάκτησης πληροφοριών. Επίσης, συντελεί με δομικό τρόπο, στη βελτίωση των διαδικασιών της ανάκτησης και της περιήγησης στα έγγραφα.



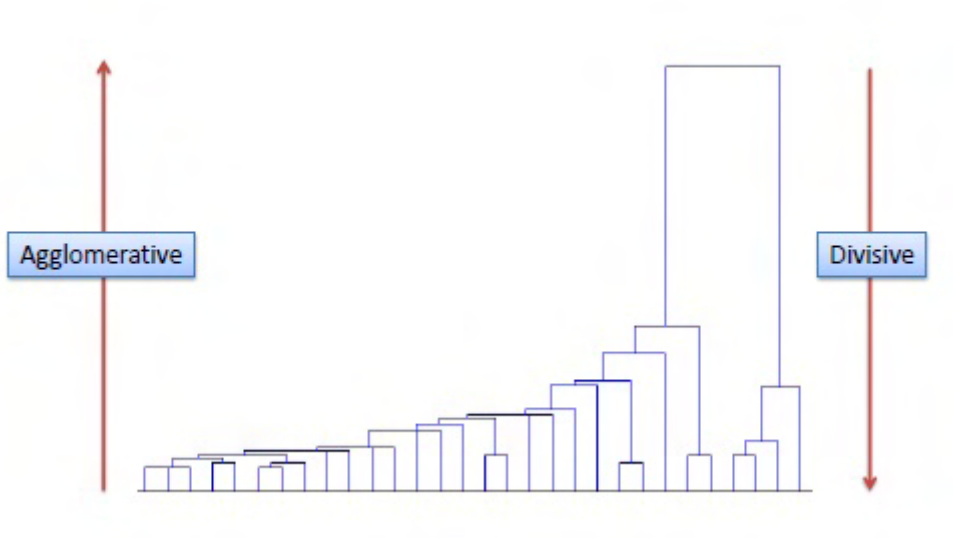
Σχήμα 4.5: Παράδειγμα Dirichlet κατανομής συναρτήσει του α (Πηγή: [30])

Οι αλγόριθμοι της ομαδοποίησης μπορούν να ταξινομηθούν σε 2 κατηγορίες, στους ιεραρχικούς και στους τμηματικούς. Οι ιεραρχικοί αλγόριθμοι, εντοπίζουν αποτελεσματικές συστάδες, αξιοποιώντας ήδη υπάρχουσες συστάδες. Βασικό βήμα για την ιεραρχική συσταδοποίηση, είναι ο καθορισμός του μέτρου απόστασης. Αρκετά συνηθισμένο παράδειγμα μέτρου απόστασης, αποτελεί η Ευκλείδια απόσταση. Επιπλέον, οι ιεραρχικοί αλγόριθμοι χωρίζονται σε 2 κατηγορίες, στους συγκεντρωτικούς (από κάτω προς τα πάνω) και στους διαιρετικούς (από πάνω προς τα κάτω). Στο Σχ. 4.6, φαίνεται η ταξινόμηση των αλγορίθμων ομαδοποίησης, μαζί με κάποια παραδείγματα. Οι συγκεντρωτικοί αλγόριθμοι συσταδοποίησης, αντιμετωπίζουν αρχικά το κάθε αντικείμενο σαν μία ολόκληρη συστάδα και συγχωνεύονται μετέπειτα σε μεγαλύτερες, ενώ οι διαιρετικοί θεωρούν όλα τα στοιχεία ως μία ολόκληρη συστάδα και στη συνέχεια τη διαιρούν σε μικρότερες. Στο Σχ. 4.7, απεικονίζεται η σχηματική διαδικασία των ιεραρχικών αλγορίθμων.

Οι τμηματικοί αλγόριθμοι, καθορίζουν εκ νέου όλες τις συστάδες την κάθε φορά, σε αντίθεση με τους ιεαραρχικούς που βασίζονται σε προηγούμενες. Βασίζονται στον προσδιορισμό ενός αρχικού αριθμού συστάδων και επαναληπτικά, επανατοποθετούν τα αντικείμενα μεταξύ των συστάδων, μέχρι ο αλγόριθμος να πετύχει σύγκλιση. Οι τμηματικοί αλγόριθμοι χωρίζονται με τη σειρά τους, σε αλγορίθμους βασισμένους στο κέντρο, στην πυκνότητα και σε φασματικούς αλγορίθμους. Στους πρώτους, το κέντρο μίας συστάδας είναι και ο μέσος όρος όλων των αντικειμένων που ανήκουν σ' αυτή και δεν αποτελεί αναγκαστικά μέρος του συνόλου των δεδομένων. Στους αλγορίθμους βασισμένους στην πυκνότητα, η δημιουργία των ομάδων γίνεται με βάση την πυκνότητα των σημείων σε μία συγκεκριμένη περιοχή. Ορίζονται περιοχές υψηλότερης πυκνότητας έναντι των υπόλοιπων σημείων, του αρχικού συνόλου δεδομένων. Η βασική προϋπόθεση, είναι

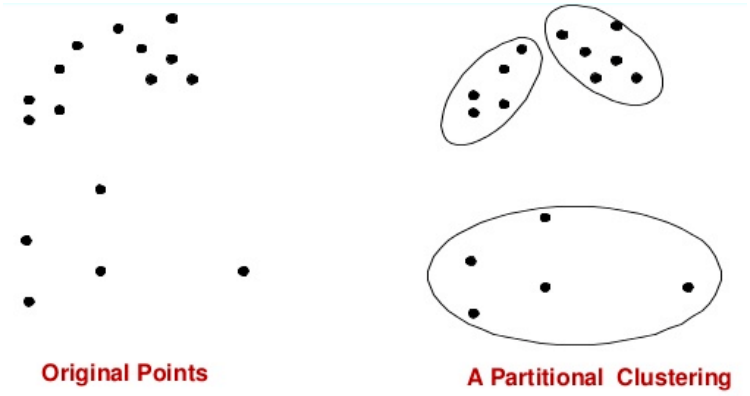


Σχήμα 4.6: Κατηγοριοποίηση των αλγορίθμων συσταδοποίησης (Πηγή: [31])



Σχήμα 4.7: Απεικόνιση συσταδοποίησης των ιεραρχικών αλγορίθμων (Πηγή: [27])

εντός της "γειτονιάς" μίας συστάδας, με συγκεκριμένη ακτίνα, να περιλαμβάνεται ένας ελάχιστος αριθμός αντικειμένων που αποτελεί και το όριο της συστάδας. Τέλος, οι φασματικοί αλγόριθμοι, διαθέτουν πολλά θεμελιώδη πλεονεκτήματα έναντι παραδοσιακών τεχνικών, όπως είναι ο αλγόριθμος K-Means. Η φασματική συσταδοποίηση, μπορεί να πραγματοποιηθεί με τυπικές μεθόδους γραμμικής άλγεβρας. Κατασκευάζει ένα γράφημα ομοιότητας για όλα τα σημεία των δεδομένων. Έπειτα, γίνεται η ένθεση των σημείων σε χώρο χαμηλότερων διαστάσεων, όπου οι συστάδες είναι πιο εμφανής, με τη χρήση ιδιοδιανυσμάτων του Laplacian γραφήματος. Ένας τυπικός αλγόριθμος, όπως ο K-Means, εφαρμόζεται τελικά, για να χωρίσει την ένθεση και να ολοκληρωθεί η διαδικασία. Στο Σχ. 4.8, δίνονται κάποια αρχικά σημεία και φαίνεται το αποτελέσμα της συσταδοποίησης τους, με τη χρήση τμηματικών αλγορίθμων.



Σχήμα 4.8: Απεικόνιση συσταδοποίησης των τμηματικών αλγορίθμων (Πηγή: [35])

4.3.2 Ο K-Means Αλγόριθμος

Ο αλγόριθμος K-Means είναι από τους πιο γνωστούς, εάν όχι ο πιο γνωστός αλγόριθμος της συσταδοποίησης. Ανήκει στους τμηματικούς αλγορίθους της συσταδοποίησης και συγκεκριμένα είναι ένας αλγόριθμος βασισμένος στο κέντρο. Ο αλγόριθμος αυτός, τοποθετεί κάθε σημείο στην κοντινότερη συστάδα, της οποίας το κέντρο ονομάζεται κεντροειδές. Το κέντρο της κάθε συστάδας, είναι το πιο αντιπροσωπευτικό σημείο της και είναι συνήθως, ο μέσος όρος όλων των τιμών των σημείων που την αποτελούν. Ο K-Means είναι "ευαίσθητος" στις ακραίες περιπτώσεις, δηλαδή σε περιπτώσεις, όπου ένα αντικείμενο έχει εξαιρετικά υψηλή τιμή και προκαλεί ουσιαστική διαστρέβλωση στην κατανομή των δεδομένων. Η διαδικασία του σχηματισμού των συστάδων με τη χρήση του αλγορίθμου K-Means, είναι η εξής:

- **Βήμα 1**. Στο πρώτο βήμα καθορίζεται ο αριθμός των ομάδων που θέλουμε να δημιουργήσει ο αλγόριθμος. Τα κέντρα των ομάδων αρχικοποιούνται τυχαία και είναι διανύσματα ίδιου μήκους, με τα διανύσματα των σημείων των δεδομένων. Σ' αυτό το στάδιο, καλό θα ήταν να υπάρχει η δυνατότητα παρατήρησης των δεδομένων και προσπάθειας εντοπισμού διαφορετικών ομάδων με "γυμνό" μάτι. Αυτό όμως, δεν είναι πολλές φορές εφικτό.

- **Βήμα 2**. Κάθε σημείο, με βάση του μέτρου της απόστασης που έχει προκαθοριστεί, υπάγεται

στη συστάδα, της οποίας το κέντρο είναι πιο κοντά στο σημείο αυτό, μετά από υπολογισμό όλων των αποστάσεων μεταξύ του σημείου και όλων των ομάδων.

- **Βήμα 3**. Σ' αυτό το στάδιο, το κέντρο της κάθε συστάδας επαναπροσδιορίζεται, με βάση το μέσο όρο όλων των διανυσμάτων των σημείων που την αποτελούν.

- **Βήμα 4**. Τα βήματα 1,2 και 3, επαναλαμβάνονται για ένα συγκεκριμένο αριθμό επαναλήψεων ή μέχρι οι αλλαγές στα κέντρα της κάθε συστάδας να είναι μηδενικές ή αμελητέες. Η εκ νέου αρχικοποίηση των κέντρων των συστάδων, προσφέρει τη δυνατότητα καινούριων εκτελέσεων, με απώτερο σκοπό την επιλογή εκείνης με τα καλύτερα αποτελέσματα.

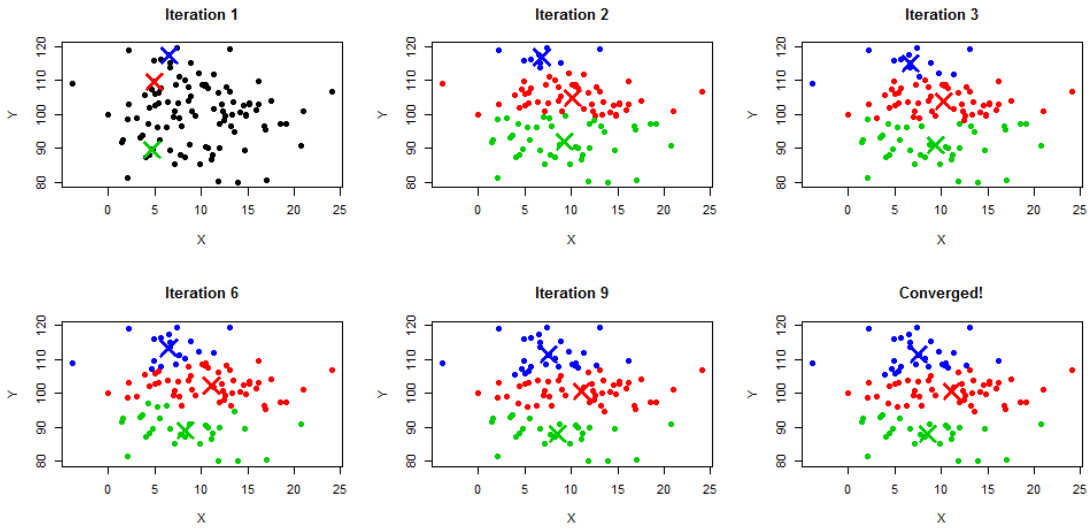
Γενικότερα, ο K-Means είναι ένας αλγόριθμος αρκετά γρήγορος και απλός στην υλοποίηση του, επειδή καθ' όλη τη διάρκεια της εκτέλεσης, οι αποστάσεις μεταξύ των κέντρων των συστάδων και των σημείων, είναι τα μοναδικά που υπολογίζονται. Διαθέτει γραμμμική πολυπλοκότητα της τάξης του O (n). Ένα από τα αρνητικά του αλγορίθμου, είναι ότι πρέπει να προκαθοριστεί ο αριθμός των συστάδων που καλείται να δημιουργήσει. Πολλές φορές, λόγω του όγκου και της πολυπλοκότητας των δεδομένων, αυτό δεν είναι εφικτό. Το ιδανικότερο θα ήταν, ο ίδιος ο αλγόριθμος να εντοπίζει τον αριθμό των αντιπροσωπευτικών ομάδων που προκύπτουν από τα δεδομένα. Τέλος, η τυχαιότητα των κέντρων των συστάδων, μπορεί να αποδίδει διαφορετικά αποτελέσματα σε κάθε επανάληψη, κάτι που οδηγεί σε μη επαναλαμβανόμενα δεδομένα με μηδενική συνοχή.

Στο Σχ. 4.9, δίνεται ένα παράδειγμα εφαρμογής του αλγορίθμου K-Means. Στην επανάληψη 1, γίνεται η αρχικοποίηση των κέντρων των συστάδων, στην επανάληψη 2, έχουμε τη νέα θέση των κέντρων, ενώ στην επανάληψη 3, καθώς κινούνται τα κέντρα, αυξάνεται ο αριθμός των μπλε σημείων. Αν μεταβούμε στην επανάληψη 6, βλέπουμε ότι το κόκκινο κέντρο έχει μετακινηθεί προς τα δεξιά, ενώ η επανάληψη 9, δείχνει ότι το πράσινο τμήμα είναι μικρότερο απ' όσο ήταν στην επανάληψη 2. Επίσης, το μπλε τμήμα έχει πάρει την κορυφή και το κόκκινο κέντρο είναι λεπτότερο σε σχέση με αυτό της επανάληψης 6. Τα αποτελέσματα της επανάληψης 9, είναι τα ίδια με τα αποτελέσματα της επανάληψης 8, οπότε ο αλγόριθμος έχει πετύχει σύγκλιση.

4.3.3 Τρόπος Καθορισμού του Κατάλληλου Αριθμού των Συστάδων

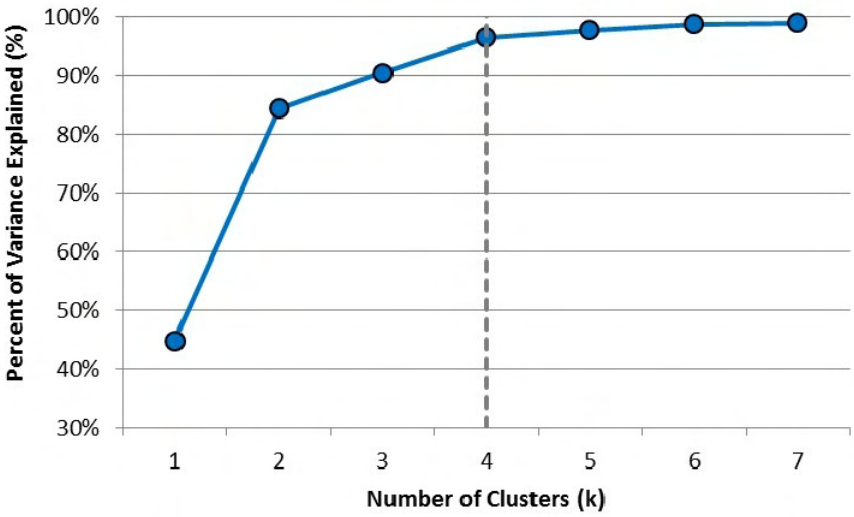
Πολλοί από τους αλγορίθμους της ομαδοποίησης, απαιτούν να είναι γνωστός ο αριθμός των συστάδων που καλούνται να δημιουργήσουν, προτού ξεκινήσει η εκτέλεσή τους. Υπάρχουν διάφορες τεχνικές, οι οποίες μπορούν να προσδιορίσουν τον κατάλληλο αριθμό των συμπλεγμάτων που μπορούν να δημιουργηθούν. Ένας απλός κανόνας, ορίζει τον αριθμό συστάδων να είναι ίσος με $k \approx \sqrt{n/2}$, όπου n, ο συνολικός αριθμός των αντικειμένων. Τα κριτήρια Elbow, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) και Deviance Information Criterion (DIC), είναι ένα σύνολο τεχνικών που συμβάλουν στη διαδικασία αυτή.

Το κριτήριο Elbow που αποτελεί μία συνηθισμένη πρακτική για την εύρεση του κατάλληλου αριθμού συστάδων, αναφέρει ότι όταν σ' ένα σύνολο από συστάδες προστεθεί μία επιπλέον συστάδα, χωρίς να προσφέρει κάποια πληροφορία, τότε ο αριθμός των συστάδων θα πρέπει να παραμείνει σταθερός, απορρίπτωντας την προσθήκη της καινούριας. Στο Σχ. 4.10, όπου ο y άξονας συμβολίζει το ποσοστό διακύμανσης που εξηγείται από τις συστάδες, δηλαδή σε ποιο βαθμό οι



Σχήμα 4.9: Παράδειγμα συσταδοποίησης με τον αλγόριθμο K-Means (Πηγή: [32])

συστάδες εκπροσωπούν το σύνολο των αντικειμένων, υπάρχει ένα σημείο όπου το οριακό κέρδος πέφτει, αποδίδοντας στο σχήμα μία γωνία. Το σημείο αυτό, δηλώνει τον καταλληλότερο αριθμό συμπλεγμάτων. Στην περίπτωση του παρακάτω σχήματος, ο καταλληλότερος αριθμός συστάδων είναι 4. Αυτός ο "αγκώνας" που δημιουργείται στο προαναφερόμενο σημείο, είναι και ο λόγος που η τεχνική αυτή ονομάστηκε Elbow. Στον άξονα y, πολλές φορές επιλέγεται να είναι η συνάρτηση Sum of Squared Errors (SSE) που υπολογίζει το άθροισμα των τετραγωνικών σφαλμάτων, τα οποία προκύπτουν από την ομαδοποίηση. Με την ίδια λογική, παρατηρούμε το σημείο που δημιουργείται ο "αγκώνας", προκειμένου να εντοπιστεί ο βέλτιστος αριθμός συστάδων.

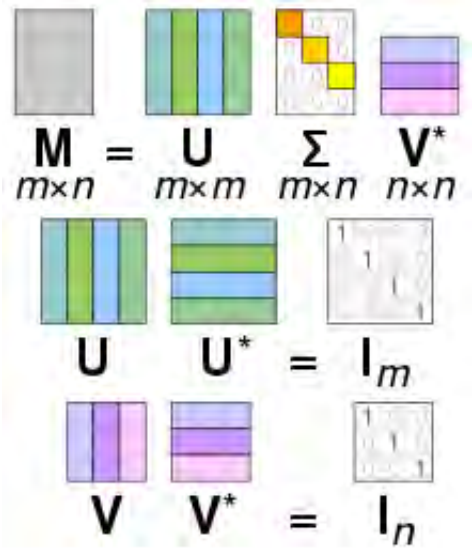


Σχήμα 4.10: Κριτήριο Elbow για τον καθορισμό του κατάλληλου αριθμού συστάδων (Πηγή: [46])

4.3.4 Αποσύνθεση Μοναδικής Τιμής

Η αποσύνθεση ενός πίνακα, γνωστή και ως παραγοντοποίηση, περιγράφει τα συστατικά του στοιχεία. Μία από τις πιο γνωστές τεχνικές παραγοντοποίησης ενός πίνακα, είναι η μέθοδος της Αποσύνθεσης Μοναδικής Τιμής. Όλοι οι πίνακες επιδέχονται αυτή τη μέθοδο και συναντάται σε πολλές εφαρμογές πινάκων, όπως είναι η συμπίεση, η μείωση των δεδομένων (διαστάσεων) ή η κατασκευή ψευδοαντίστροφου πίνακα. Η τεχνική αυτή, απλοποιεί ένα δοσμένο πίνακα, μειώνοντας τις διαστάσεις του, περιλαμβάνοντας μόνο τα συστατικά του στοιχεία, με απώτερο σκοπό να καταστήσει πιο απλούς, τους υπολογισμούς που πρόκειται να ακολουθήσουν. Ο πίνακας μπορεί να περιλαμβάνει είτε πραγματικές τιμές, είτε μιγαδικές. Στο δικό μας μοντέλο, ο πίνακας που αποσυντέθηκε, αφορά πραγματικές τιμές.

Έστω έχουμε έναν πίνακα M με πραγματικές ή μιγαδικές τιμές, διαστάσεων m x n που θέλουμε να αποσυνθέσουμε. Η διαδικασία αποσύνθεσης περιγράφεται από τον μαθηματικό τύπο $M = U\Sigma V^T$, όπου U είναι ένας πραγματικός ή μιγαδικός πίνακας διαστάσεων m x m, ο Σ είναι ένας ορθογώνιος διαγώνιος πίνακας διαστάσεων m x n, με μη αρνητικές πραγματικές τιμές στη διαγώνιό του και ο V^T, είναι ο ανάστροφος ενός πίνακα V, διαστάσεων n x n, όπου ο V είναι ένας μοναδιαίος πραγματικός ή μιγαδικός πίνακας. Στον πίνακα Σ, οι καταχωρήσεις στη διαγώνιο, είναι και οι μοναδικές τιμές του M, ενώ οι στήλες του U και του V, εκπροσωπούν τα αριστερά και τα δεξιά μοναδικά διανύσματα του M αντίστοιχα. Τα αριστερά μοναδικά διανύσματα του M, είναι ένα σύνολο από ορθοκανονικά ιδιοδιανύσματα του MM^T, τα δεξιά είναι ένα σύνολο απο ορθοκανονικά ιδιοδιανύσματα του M^TM και τα στοιχεία στη διαγώνιο του Σ, είναι οι τετραγωνικές ρίζες των μη μηδενικών ιδιοτιμών του M^TM και του MM^T. Στο Σχ. 4.11, απεικονίζεται η διαδικασία της παραγοντοποίησης ενός M πίνακα, με διαστάσεις m x n.



Σχήμα 4.11: Απεικόνιση αποσύνθεσης μοναδικής τιμής ενός πίνακα M (Πηγή: `https://en.wikipedia.org/wiki/Singular_value_decomposition`)

4.4 Πρόβλεψη Μελλοντικών Αναφορών

4.4.1 Εισαγωγή

Η αξιολόγηση του παρελθόντος ενός επιστήμονα και η πρόβλεψη του μελλοντικού του αντίκτυπου, αποτελούν τη βάση για τη λήψη αποφάσεων σχετικά με τις προσλήψεις και τις χρηματοδοτήσεις που λαμβάνει. Η αναγνωρισιμότητα και η επιτυχία του, συνδέονται όλο και περισσότερο με τον αριθμό των παραπομπών που λαμβάνουν οι δημοσιεύσεις του. Παρ' όλα αυτά, δεν είναι λίγες οι φορές που οι ερευνητές αναφέρουν μόνο ένα μικρό ποσοστό ερευνών/εργασιών που τους ενδιαφέρει. Ο έγκυρος εντοπισμός μίας πολλά υποσχόμενης εργασίας ή έρευνας, προτού λάβει αρκετές αναφορές, είναι χρήσιμος από πολλές απόψεις, τόσο για τους ίδιους τους συγγραφείς όσο και για την ευρεία επιστημονική κοινότητα.

Τις περισσότερες φορές, οι επιστήμονες αξιολογούνται με βάση την προηγούμενη βιβλιογραφία τους, η οποία σχετίζεται με τους αντίστοιχους ερευνητικούς τομείς. Ωστόσο, αυτό δεν είναι πάντα αποδεκτό ή εφικτό, καθώς ο όγκος των δημοσιευμένων ερευνών αυξάνεται εκθετικά. Δεδομένου όμως αυτού του όγκου των δημοσιεύσεων, είναι αξιοσημείωτο να μπορούμε να προβλέπουμε τις μελλοντικές αναφορές που θα έχει ένα έγγραφο, με βάση το πλήθος των αναφορών του τα προηγούμενα χρόνια. Η καταγραφή όλων αυτών των παραπομπών των προηγούμενων ετών, αποτυπώνεται σε μία ακολουθία αριθμών (χρονοσειρά), η οποία αποτελεί τη βάση για την εκπαίδευση του νευρωνικού δικτύου που θα πραγματοποιήσει τις προβλέψεις.

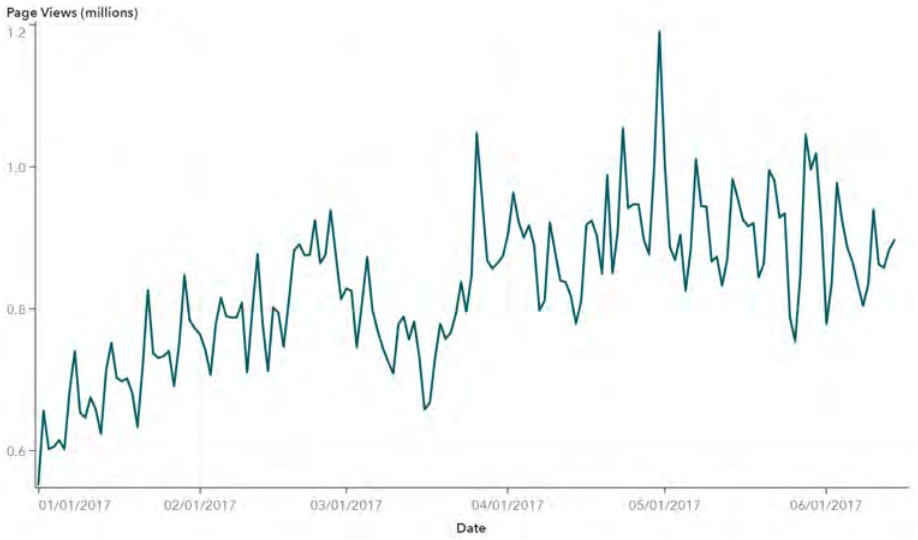
4.4.2 Χρονολογικές Σειρές

Με τον όρο χρονολογική σειρά, αναφερόμαστε σε μία σειρά από παρατηρήσεις που γίνονται σε συγκεκριμένες χρονικές στιγμές ή περιόδους, με κύριο χαρακτηριστικό ότι αυτές οι στιμές ή οι περίοδοι, ισαπέχουν μεταξύ τους. Το διάστημα μεταξύ αυτών, θα πρέπει να είναι ισόποσο και όχι ανόμοιο. Σε διαφορετική περίπτωση, δεν μπορούμε να αναφερόμαστε στον όρο χρονολογική σειρά. Κάθε παρατήρηση αποτελεί και μία τυχαία μεταβλητή και αυτός είναι και ο λόγος που η διαδικασία που αφορά τις χρονολογικές σειρές, είναι στοχαστική. Τέτοιες χρονολογικές σειρές μπορούμε να εντοπίσουμε σε διάφορα επιστημονικά πεδία, όπως στο πεδίο των οικονομικών (χρηματιστήριο), των ιατρικών (παλμογράφοι) κ.λπ. Στις χρονοσειρές θα πρέπει να εντοπίζεται πάντα το στοιχείο της αλληλεξάρτησης μεταξύ των παρατηρήσεων, με απώτερο σκοπό τη δημιουργία ανταγωνιστικών μοντέλων πρόβλεψης, με τα καλύτερα δυνατά αποτελέσματα. Στο Σχ. 4.12, απεικονίζεται ένα παράδειγμα μίας χρονολογικής σειράς που αφορά την ημερήσια επισκεψιμότητα μίας ιστοσελίδας, από τους χρήστες του διαδικτύου.

Τα δεδομένα των χρονολογικών σειρών, μπορούν να χαρακτηριστούν ως διαδικασία εποπτευόμενης μάθησης. Για ένα σύνολο δεδομένων χρονοσειρών, μπορούμε να αναδιαρθρώσουμε τα δεδομένα, προκειμένου να αποτελεί πρόβλημα εποπτευόμενης μάθησης. Αυτό καθίσταται εφικτό, εάν χρησιμοποιήσουμε τα προηγούμενα βήματα χρόνου ως μεταβλητές εισόδου και τα επόμενα βήματα χρόνου ως μεταβλητές εξόδου, ώστε να υπάρχει η λειτουργία χαρτογράφησης από την είσοδο στην έξοδο. Έτσι, η διαδικασία επεξεργασίας δεδομένων από ένα σύνολο χρονολογικών σειρών, μπορεί να χαρακτηριστεί ως εποπτευόμενη μάθηση.

Μία χρονολογική σειρά μπορεί να διέπεται από ένα ή και περισσότερα χαρακτηριστικά. Αυτά τα χαρακτηριστικά είναι τα εξής:

- **Τάση**. Με τον όρο τάση, αναφερόμαστε στη συνολική πορεία που χαρακτηρίζει την χρονοσειρά. Η χρονοσειρά μπορεί να χαρακτηριστεί ως αύξουσα, στη περίπτωση που οι τιμές αυξάνονται συναρτήσει του χρόνου ή φθίνουσα, όταν οι τιμές μειώνονται.
- **Εποχιακή Συνιστώσα ή Εποχικότητα**. Ταυτίζεται σε μεγάλο βαθμό με το χαρακτηριστικό της τάσης, με διαφορά ότι πρέπει για να μπορεί να χρησιμοποιηθεί ο παραπάνω όρος, η χρονοσειρά να επαναλαμβάνεται περιοδικά σε συνάρτηση με το χρόνο (εβδομαδιαία, μηνιαία, κ.λπ). Η τάση και η εποχιακή συνιστώσα είναι 2 χαρακτηριστικά που μπορούν να διέπουν τη χρονοσειρά ταυτόχρονα.
- **Τυχαία Συνιστώσα**. Οι μεταβολές που συναντώνται στην περίπτωση αυτή, δεν υπάγονται στις 2 προαναφερόμενες κατηγορίες και σχετίζονται με μεταβολές που οφείλονται στην ύπαρξη θορύβου ή άλλων τυχαίων γεγονότων.



Σχήμα 4.12: Παράδειγμα χρονολογικής σειράς για το πλήθος των χρηστών που επισκέπτεται μία ιστοσελίδα (Πηγή: [57])

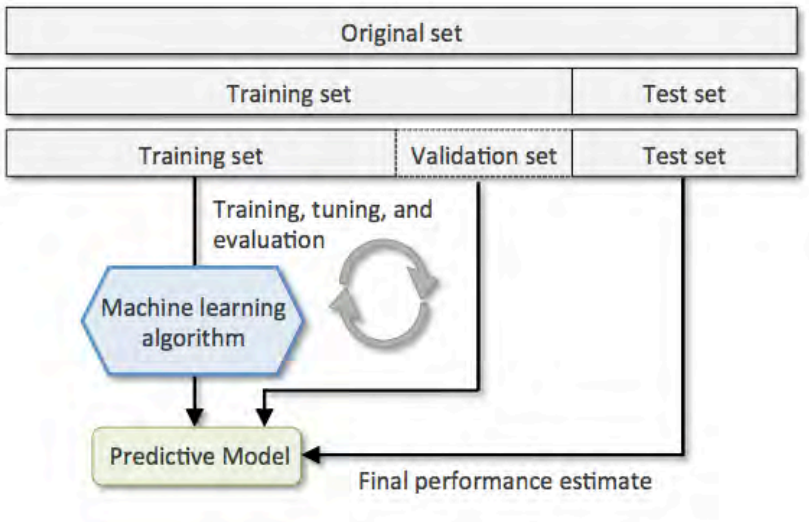
4.4.3 Δεδομένα Εκπαίδευσης, Επικύρωσης και Δοκιμής

Στην ανάπτυξη και στην εκπαίδευση μοντέλων, συντελεί σε μεγάλο βαθμό, ο διαχωρισμός των αρχικών δεδομένων σε διάφορα υποσύνολα. Πολλές τεχνικές έχουν αναπτυχθεί σχετικά με τον διαχωρισμό του αρχικού συνόλου. Τα ποσοστά διαχωρισμού των δεδομένων της παρακάτω διαδικασίας, δεν είναι σταθερά, αλλά εξαρτώνται κυρίως από το συνολικό αριθμό δειγμάτων που έχουμε στην διάθεση μας και από το πραγματικό μοντέλο που πρόκειται να εκπαιδευτεί. Στο Σχ. 4.13, απεικονίζεται η διαίρεση του αρχικού συνόλου των δεδομένων. Συγκεκριμένα, 3 είναι τα σύνολα που χρησιμοποιούνται στα διάφορα στάδια της εκπαίδευσης του μοντέλου και είναι τα εξής:

- **Δεδομένα Εκπαίδευσης**. Αποτελεί το μεγαλύτερο σετ δεδομένων από τα 3, περίπου το 70% και χρησιμοποιείται για την εκπαίδευση του αλγορίθμου. Το μοντέλο "βλέπει" και μαθαίνει απ' αυτά τα δεδομένα. Ένα μέρος του συνόλου εκπαίδευσης, χρησιμοποιείται για να δημιουργηθούν τα δεδομένα επικύρωσης.

- **Δεδομένα Επικύρωσης**. Είναι το σύνολο των δεδομένων που παρέχει μία αμερόληπτη αξιολόγηση ενός μοντέλου (όχι του τελικού), προσαρμοσμένο στα δεδομένα εκπαίδευσης, ρυθμίζοντας παράλληλα τις υπερ-παραμέτρους του μοντέλου που αναφέραμε και στο 2.3.2. Αποτελεί συνήθως το 10%.

- **Δεδομένα Δοκιμής**. Αυτό το σύνολο δεδομένων αξιοποιείται μόνο μία φορά, όταν η εκπαίδευση του μοντέλου έχει ολοκληρωθεί. Παρέχει αξιολόγηση του τελικού μοντέλου που προσαρμόστηκε στα δεδομένα εκπαίδευσης και επικύρωσης και αποτελεί συνήθως το 20%.

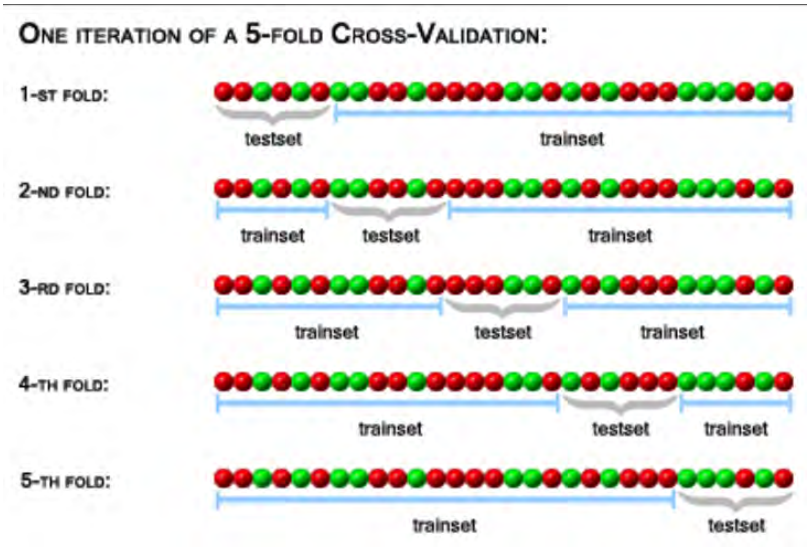


Σχήμα 4.13: Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης, επικύρωσης και δοκιμής (Πηγή: [28])

,

4.4.4 K-fold Cross Validation

Για τη βελτιστοποίηση της εκπαίδευσης ενός μοντέλου, λόγω της τυχαιότητας που χαρακτηρίζει τη διαίρεση των αρχικών δεδομένων, δίνεται η δυνατότητα να αξιοποιηθεί η μέθοδος K-fold Cross Validation, η οποία είναι μία διαδικασία επαναδειγματοληψίας που χρησιμοποιείται για την αξιολόγηση μοντέλων της M.M, σ' ένα περιορισμένο δείγμα δεδομένων. Έχοντας αρχικά τα δεδομένα, ανάλογα με την τιμή της παραμέτρου K που ορίζεται, κατασκευάζει τόσα σύνολα δεδομένων, αποτελούμενα από δεδομένα εκπαίδευσης και δοκιμής, χωρισμένα με τυχαίο τρόπο. Έτσι, αν δώσουμε την τιμή 5 στην παράμετρο K, τότε θα δημιουργηθούν 5 σύνολα δεδομένων, όπου το κάθε σύνολο είναι χωρισμένο τυχαία σε δεδομένα εκπαίδευσης και δοκιμής. Στο Σχ. 4.14, απεικονίζεται η μέθοδος K-fold Cross Validation που περιγράψαμε.



Σχήμα 4.14: Μία επανάληψη της μεθόδου K-fold Cross Validation (Πηγή: [15])

,

4.4.5 Το φαινόμενο της Υπερ-προσαρμογής και της Υπο-προσαρμογής

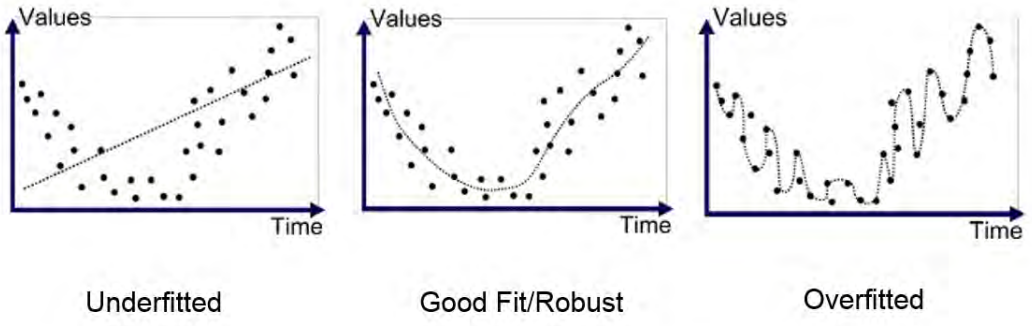
Σε στατιστικά μοντέλα ή σε μοντέλα της Μ.Μ, όπως τα μοντέλα ταξινόμησης ή πρόβλεψης, είναι συχνό να συναντάται το φαινόμενο της υπερ-προσαρμογής ή της υπο-προσαρμογής, το οποίο σχετίζεται με τα δεδομένα εκπαίδευσης του εκάστοτε μοντέλου. Όταν δεν συναντάται κανένα απ' αυτά τα φαινόμενα, το μοντέλο πετυχαίνει καλά αποτελέσματα και δεν χρειάζεται κάποια περαιτέρω επεξεργασία στα δεδομένα εκπαίδευσης. Συχνά όμως, τα 2 φαινόμενα αποτελούν μέρος του προβλήματος, δημιουργώντας δυσχερές συνθήκες στην προσπάθεια εκπαίδευσης του μοντέλου.

Η υπερ-προσαρμογή, χαρακτηρίζει ένα μοντέλο, όταν έχει μάθει πολύ καλά τα δεδομένα εκπαίδευσης και καταγράφει συγχρόνως, τους θορύβους που τα συνοδεύουν, με αποτελέσμα το μοντέλο να μην είναι αποτελεσματικό σε δεδομένα που δεν έχει "ξανασυναντήσει" (δεδομένα δοκιμής). Συγκεκριμένα, η υπερ-προσαρμογή εμφανίζεται όταν υπάρχει χαμηλή μεροληψία αλλά μεγάλη διακύμανση. Ο κυριότερος λόγος που μπορεί να συμβεί αυτό, είναι ο αναποτελεσματικός διαχωρισμός του αρχικού συνόλου δεδομένων σε δεδομένα εκπαίδευσης, επικύρωσης και δοκιμής. Η μεροληψία αντιπροσωπεύει την τάση του αλγορίθμου να μαθαίνει το λάθος με συστηματικό τρόπο, χωρίς να λαμβάνει όλες τις πληροφορίες που απαρτίζουν τα δεδομένα, ενώ η διακύμανση, αντιπροσωπεύει την "ευαισθησία" του αλγορίθμου στους θορύβους, με αποτελέσμα να μην μπορεί να λάβει το πραγματικό σήμα από το σύνολο δεδομένων. Αντιμετωπίζεται με εκτέλεση πλήθους δοκιμών διαχωρισμού στα αρχικά δεδομένα.

Από την άλλη, η υπο-προσαρμογή διέπει ένα μοντέλο, όταν το μοντέλο δεν μπορεί να μάθει την τάση που χαρακτηρίζει τα δεδομένα εκπαίδευσης και ο κυριότερος λόγος είναι η ύπαρξη ανεπαρκών δεδομένων εκπαίδευσης. Αναλυτικότερα, η υπο-προσαρμογή εμφανίζεται όταν το μοντέλο εμφανίζει χαμηλή διακύμανση αλλά υψηλή μεροληψία. Συνήθως, τα μοντέλα που χαρακτηρίζονται από υπο-προσαρμογή, είναι και εκείνα που δομήθηκαν πάνω σε πολύ απλές προϋποθέσεις. Η προσθήκη περισσότερων δεδομένων στο αρχικό σύνολο, αποτελεί μία από τις κυριότερες λύσεις

για την αντιμετώπιση του φαινομένου.

Γενικότερα, είτε συναντάται το φαινόμενο της υπερ-προμοσαρμογής είτε της υπο-προσαρμογής, το μοντέλο είναι ανίκανο να κάνει σωστές προβέψεις σε καινούρια σύνολα δεδομένων και ένας τρόπος για την εξάλειψη και των 2 φαινομένων, είναι η χρήση δεδομένων επικύρωσης. Στο Σχ. 4.15, παρουσιάζονται τα 2 φαινόμενα και απεικονίζεται η περίπτωση, όπου το μοντέλο εκπαιδεύεται σωστά από τα δεδομένα εκπαίδευσης.



Σχήμα 4.15: Παράδειγμα υπερ-προσαρμογής και υπο-προσαρμογής (Πηγή: [64])

4.4.6 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούν σχετικά μία καινούρια περιοχή μελέτης, η οποία ερευνάται τα τελευταία 40 χρόνια. Παρ' όλα αυτά, λόγω της ραγδαίας αύξησης της ζήτησής τους και της μεγάλης εφαρμογής τους στην ανάπτυξη μοντέλων και ειδκότερα σε μοντέλα πρόβλεψης, έχουν επιφέρει μεγάλη επανάσταση στον τομέα της Μ.Μ. Ιδιαίτερα δημοφιλή ξεκίνησαν να είναι το 1990. Τα νευρωνικά δίκτυα υπάγονται στη Β.Μ που αποτελεί υποπεδίο της Μ.Μ και μπορούν να εφαρμοστούν σχεδόν σε κάθε περίπτωση, όπου συναντάται η ύπαρξη εξαρτημένων και ανεξάρτητων μεταβλητών.

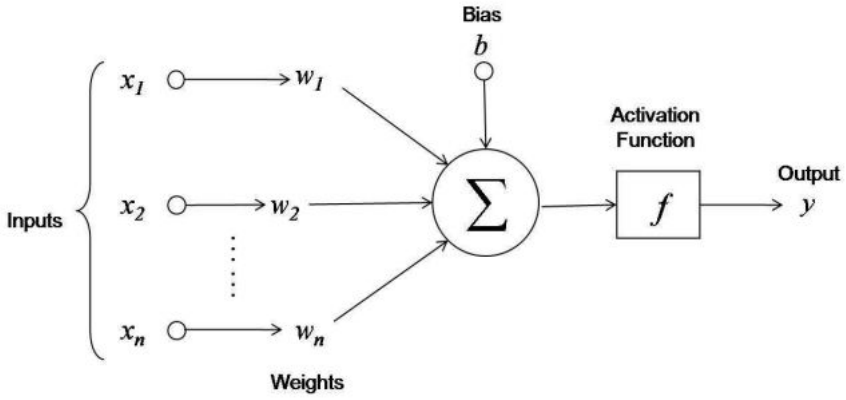
Η δομή των νευρωνικών δικτύων στηρίζεται στα βιολογικά νευρωνικά δίκτυα που διαθέτει ο άνθρωπος. Τα σημερινά αποκαλούμενα "βαθιά" νευρωνικά δίκτυα, έχουν αποδειχθεί ότι δουλεύουν αρκετά καλά. Ουσιαστικά, τα νευρωνικά δίκτυα είναι προσεγγίσεις σε συναρτήσεις και αυτός είναι και ο λόγος που μπορούν να αξιοποιηθούν κάθε φορά που υφίσταται το πρόβλημα της πολύπλοκης χαρτογράφησης, από το χώρο των εισόδων στο χώρο των εξόδων. Λαμβάνουν ερεθίσματα από την είσοδο, εκπαιδεύονται και συμπεριφέρονται αναλόγως. Αξίζει να σημειωθεί, ότι τα νευρωνικά δίκτυα χρησιμοποιούνται για προβλήματα ταξινόμησης και παλινδρόμησης. Σκοπός της πρώτης είναι η ταξινόμηση όμοιων αντικειμένων στην ίδια ομάδα και σκοπός της τελευταίας, είναι η δημιουργία μίας ευθείας γραμμής στο δισδιάστατο χώρο, με σκοπό την ελαχιστοποίηση των σφαλμάτων μεταξύ των πραγματικών και των προβλεπόμενων τιμών, δηλαδή των τιμών αυτών που εμφανίζονται σε μία σχηματιζόμενη ευθεία.

Στο εσωτερικό των νευρωνικών δικτύων, υπάρχει ένα διασυνδεδεμένο δίκτυο που αποτελείται από υπολογιστικούς κόμβους (νευρώνες). Υπάρχουν 3 τύποι νευρώνων. Οι πρώτοι είναι οι νευρώνες εισόδου, ακολουθούν οι υπολογιστικοί νευρώνες ή οι κρυμμένοι νευρώνες και οι νευ-

ρώνες εξόδου. Το νευρωνικό μέσω των νευρώνων εισόδου, λαμβάνει τα αρχικά δεδομένα ή τα αποτελέσματα προηγούμενου δικτύου και ο αριθμός των νευρώνων εισόδου ισούται τις περισσότερες φορές, με τον αριθμό των χαρακτηριστικών που διέπουν τα δεδομένα. Στόχος τους η προώθηση των δεδομένων στο επόμενο επίπεδο που απαρτίζεται από τους κρυμμένους νευρώνες. Οι τελευταίοι, λαμβάνουν αυτά τα δεδομένα και είναι υπεύθυνοι για την εκτέλεση κάποιων μετασχηματισμών πάνω στα δεδομένα. Συγκεκριμένα, υπάρχουν κάποιες ακμές που χαρακτηρίζονται από κάποια βάρη, τα οποία πολλαπλασιάζονται με τα δεδομένα εισόδου και έπειτα αθροίζονται όλα μαζί (σταθμισμένο άθροισμα των εισόδων), για να προκύψει τελικά η λεγόμενη καθαρή είσοδος. Αν υφίσταται μεροληψία, αθροίζεται και αυτή. Η καθαρή είσοδος, είναι ένας αριθμός ή πίνακας αριθμών, με τον αριθμό των γραμμών να ισούται με τον αριθμό των νευρώνων σε κάθε επίπεδο και τον αριθμό των στηλών να ισούται με τον αριθμό των δεδομένων εισόδου. Αυτό εξαρτάται με το αν αναφερόμαστε σε νευρωνικό πολλών νευρώνων με ένα ή πολλά επίπεδα ή και τα 2 μαζί ή για δίκτυο με ένα νευρώνα και μία είσοδο. Για τον καθορισμό του κατάλληλου αριθμού νευρώνων, υπάρχουν πολλές αναφορές και μεθοδολογίες, συνίσταται όμως η εκτέλεση αρκετών δοκιμών. Τέλος, εφαρμόζεται η συνάρτηση ενεργοποίησης ή αλλιώς συνάρτηση μεταφοράς που έχει προκαθοριστεί και ενεργοποιείται εφόσον το αποτέλεσμα της πρόσθεσης ξεπεράσει ένα προκαθορισμένο όριο. Το κάθε δίκτυο, μπορεί να έχει μόνο μία συνάρτηση ενεργοποίησης που εφαρμόζεται σε όλους τους νευρώνες του δικτύου και το γεγονός ότι διαθέτει τέτοια συνάρτηση, το καθιστά ικανό να εντοπίζει τις μη γραμμικές σχέσεις μεταξύ των δεδομένων. Έτσι, παράγεται η έξοδος του νευρωνικού δικτύου, η οποία θα αξιοποιηθεί στο μοντέλο της παλινδρόμησης ή στο μοντέλο της ταξινόμησης. Τα νευρωνικά δίκτυα απαιτούν μεγάλο όγκο αριθμητικών δεδομένων και η εκπαίδευσή τους μπορεί να είναι αρκετά χρονοβόρα.

Στο Σχ. 4.16, δίνεται ένα παράδειγμα ενός απλού κόμβου μέσα σ' ένα νευρωνικό δίκτυο. Εάν παρομοιάσουμε τον παρακάτω κόμβο με έναν βιολογικό νευρώνα, το βάρος αντιστοιχεί στην ισχύ μίας συνάψεως, το άθροισμα και η συνάρτηση μεταφοράς αντιπροσωπεύει το σώμα του κυττάρου και η έξοδος του νευρώνα αντιπροσωπεύει το σήμα στον άξονα.

$$\Sigma = (x_1 * w_1 + x_2 * w_2 + ... + x_n * w_n + b)$$



Σχήμα 4.16: Η δομή ενός κόμβου μέσα στο νευρωνικό δίκτυο (Πηγή: [8])

Παρακάτω, διευκρινίζονται τα βασικά βήματα για την εξαγωγή της τελικής εξόδου του νευρωνικού δικτύου, τα οποία είναι τα εξής:

- Συλλογή των δεδομένων και διαχωρισμός σε δεδομένα εκπαίδευσης, επικύρωσης και δοκιμής.
- Ανάπτυξη της δομής του δικτύου.
- Κατάλληλη επιλογή του αλγορίθμου εκπαίδευσης του δικτύου.
- Αρχικοποίηση των παραμέτρων του αλγορίθμου και των βαρών των ακμών.
- Προετοιμασία δεδομένων.
- Έναρξη εκπαίδευσης του δικτύου.
- Έλεγχος και εκτίμηση του δικτύου με χρήση μέτρων αξιολόγησης.
- Χρήση του τελικού νευρωνικού δικτύου. Εάν δεν είμαστε ικανοποιημένοι από μία φάση, επιβάλλεται να επιστρέψουμε στην προηγούμενη, για να μπορέσει το δίκτυο να έχει μία ικανοποιητική πορεία.

4.4.7 Συναρτήσεις Ενεργοποίησης

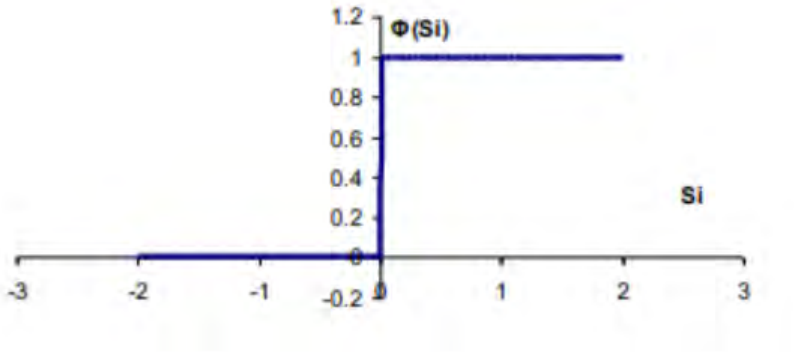
Όπως αναφέραμε προηγουμένως, τα δεδομένα προτού αποτελέσουν την τελική έξοδο του νευρωνικού δικτύου, "εισέρχονται" σε μία συνάρτηση ενεργοποίησης. Η συνάρτηση αυτή, μπορεί να είναι είτε γραμμική είτε μη γραμμική. Στα παρακάτω σχήματα των συναρτήσεων ενεργοποίησης, με S_i συμβολίζουμε το σταθμισμένο άθροισμα. Υπάρχουν πολλές τέτοιες συναρτήσεις ενεργοποίησης και οι πιο ευρέως γνωστές είναι:

- **Βηματική συνάρτηση**. Με την βηματική συνάρτηση, όπως φαίνεται και στο Σχ. 4.17, το σταθμισμένο άθροισμα αν είναι μεγαλύτερο του 0, τότε η συνάρτηση μετατρέπει το αποτέλεσμα σε 1, ενώ αν είναι μικρότερο ή ίσο με το 0, μετατρέπει το αποτελέσμα σε 0. Στη συγκεκριμένη περίπτωση, το όριο είναι το 0. Ο μαθηματικός τύπος της βηματικής συνάρτησης είναι ο εξής:

$$\Phi(S_i) = \begin{cases} 1, & \text{εάν } S_i > 0 \\ 0, & \text{εάν } Si \leq 0 \end{cases}$$

- **Γραμμική συνάρτηση**. Χρησιμοποιείται τις περισσότερες φορές για την επίλυση προβλημάτων γραμμικής παλινδρόμησης. Στο Σχ. 4.18, δίνεται η γραφική απεικόνιση της γραμμικής συνάρτησης. Ο μαθηματικός τύπος της γραμμικής συνάρτησης είναι ο εξής:

$$\Phi(S_i) = \lambda S_i$$



Σχήμα 4.17: Βηματική συνάρτηση (Πηγή: [72])

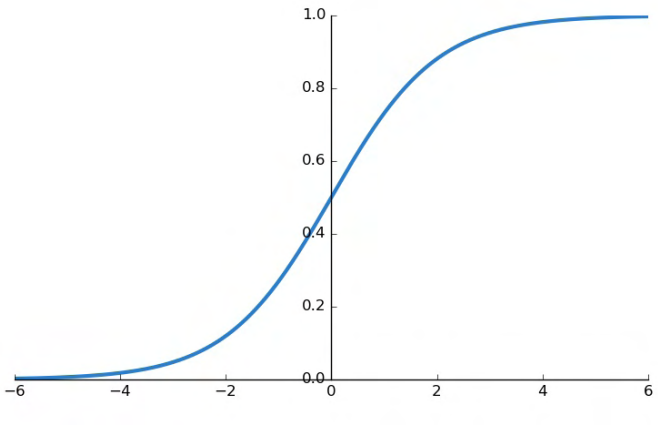
Σχήμα 4.18: Γραμμική συνάρτηση (Πηγή: [72])

- **Σιγμοειδής συνάρτηση ή σιγμοειδής καμπύλη**. Αυτή η συνάρτηση, χρησιμοποιείται για μη γραμμικά προβλήματα και συναντάται πολλές φορές, σε νευρωνικά δίκτυα πολλών επιπέδων. Στο Σχ. 4.19, απεικονίζεται η παραπάνω συνάρτηση. Ο μαθηματικός τύπος της είναι ο εξής:

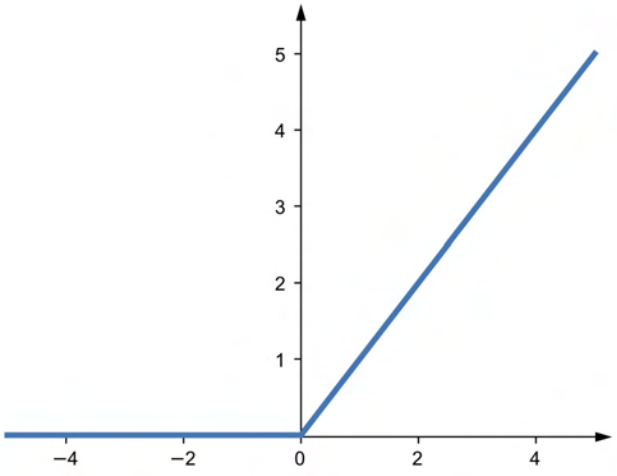
$$\Phi(S_i) = \frac{1}{1+e^{-\alpha S_i}}$$

- **Διορθωμένη Γραμμική Μονάδα (ReLU)**. Είναι από τις πιο χρησιμοποιούμενες συναρτήσεις ενεργοποίησης και εφαρμόστηκε στο νευρωνικό πρόβλεψης αναφορών που αναπτύχθηκε στην παρούσα εργασία. Συναντάται πολύ συχνά στα συνελικτικά νευρωνικά δίκτυα, τα οποία χρησιμοποιούνται κυρίως για αναγνώριση εικόνας και βίντεο. Στο Σχ. 4.20, δίνεται η γραφική απεικόνιση της συνάρτησης. Ο μαθηματικός τύπος της είναι ο εξής:

$$\Phi(S_i) = \begin{cases} \lambda S_i, & \text{εάν } S_i > 0 \\ 0, & \text{εάν } Si \leq 0 \end{cases}$$



Σχήμα 4.19: Σιγμοειδής συνάρτηση (Πηγή: [21])



Σχήμα 4.20: Συνάρτηση διορθωμένης γραμμικής μονάδας (Πηγή: [65])

4.4.8 Τύποι Νευρωνικών Δικτύων

Ο αριθμός των τύπων των νευρωνικών δικτύων, αυξάνεται εκθετικά και υπάρχουν ποικίλες αρχιτετκτονικές και προσεγγίσεις. Κάποιοι από τους βασικότερους τύπους είναι οι εξής:

- **Feed Forward (FF)**. Είναι παλιός τύπος νευρωνικού δικτύου και διαθέτει απλή μορφή και λειτουργία. Υπάρχουν δίκτυα τύπου FF, με μόνο ένα επίπεδο μεταξύ του επιπέδου εισόδου και εξόδου ή και με περισσότερα. Όλοι οι κόμβοι είναι πλήρως συνδεδεμένοι, δηλαδή κάθε νευρώνας σ' ένα επίπεδο, είναι διασυνδεμένος με όλους τους νευρώνες του προηγούμενου επιπέδου και ο κάθε κόμβος τροφοδοτείται από την έξοδο του προηγούμενου επιπέδου. Διαθέτει βρόχους που μεταφέρουν τις πληροφορίες μόνο από την είσοδο στην έξοδο, χωρίς να μπορεί να συμβεί το αντίθετο. Στις περισσότερες περιπτώσεις εκπαίδευσης αυτών των τύπων των δικτύων, χρησιμοποιείται ο αλγόριθμος Error Back Propagation που είναι μία μέθοδος για τις ενημερώσεις των βαρών του δικτύου και λειτουργεί επαναληπτικά μέχρι να παραχθεί η επιθυμητή έξοδος. Αυτό πραγματοποιείται με την κατάλληλη ρύθμιση των βα-

ρών. Η συνάρτηση που χρησιμοποιείται ως συνάρτηση ενεργοποίησης, είναι μη γραμμική και είθισται να είναι η σιγμοειδής συνάρτηση. Λόγω των συνεχών τιμών που παράγει η παραπάνω συνάρτηση, είναι εφικτή η παραγώγιση της, η οποία συνδέεται με την ανανέωση των βαρών του δικτύου.

Ο αλγόριθμος Error Back Propagation, είναι μία τεχνική, η οποία ακολουθεί τη μέθοδο της καθόδου κλίσης και εκμεταλλεύεται τον κανόνα της αλυσίδας. Όταν όλα τα στοιχεία θα βρίσκονται στην έξοδο, τότε μπορούμε να υπολογίσουμε το σφάλμα του μοντέλου, αφαιρώντας τις τιμές που έχει παράγει το νευρωνικό από τις πραγματικές (επιθυμητές) τιμές, χρησιμοποιώντας κάποια συνάρτηση κόστους. Κάθε βάρος έχει τη δική του συμβολή στο σφάλμα που εντοπίζεται με τις μερικές παραγώγους του τελικού κόμβου και στη συνέχεια εφαρμόζεται ο αλγόριθμος της καθόδου κλίσης. Έπειτα, η καινούρια πληροφορία μεταφέρεται στο προηγούμενο επίπεδο, όπου η διαδικασία υπολογισμού μερικών παραγώγων πραγματοποιείται ξανά, με τη διαφορά ότι αξιοποιείται και η προηγούμενη πληροφορία για τα βάρη, με αποτέλεσμα να επαναπροσδιορίζεται η τιμή τους. Η διαδικασία τερματίζει στο επίπεδο εισόδου, το οποίο παραμένει αμετάβλητο και αυτό γίνεται όταν φτάσει τον ζητούμενο αριμό επαναλήψεων ή όταν το σφάλμα πέσει κάτω από κάποιο όριο. Οι 2 τελευταίοι παράμετροι προκαθορίζονται από τον ίδιο τον χρήστη. Η μορφή της συνάρτησης που χρησιμοποιείται για την ανανέωση των βαρών είναι η εξής:

$$w = w - \eta \nabla_w J_w, \tag{4.1}$$

όπου J μία συνάρτηση κόστους και η, ο ρυθμός μάθησης που ανανεώνονται τα βάρη. Ο ρυθμός μάθησης μπορεί να είναι σταθερός ή να αλλάζει προσαρμοστικά κατα τη διάρκεια της εκπαίδευσης. Η παράγωγος της συνάρτησης κόστους σε σχέση με το βάρος, χρησιμοποιώντας τον κανόνα της αλυσίδας, έχει την ακόλουθη μορφή:

$$\frac{\partial J(n,y)}{\partial w} = \frac{\partial J(n,y)}{\partial \alpha} \frac{\partial \alpha}{\partial n} \frac{\partial n}{\partial w} \tag{4.2}$$

όπου n η καθαρή είσοδος. Έτσι, η αρχική Εξίσωση 4.1 ανανέωσης βαρών γίνεται:

$$w = w - \eta \frac{\partial J(n,y)}{\partial w} \tag{4.3}$$

- **Auto Encoder (AE)**. Συνηθίζεται να χρησιμοποιείται για ταξινόμηση, συσταδοποίηση και για συμπίεση χαρακτηριστικών. Στα FF νευρωνικά δίκτυα, πραγματοποιείται επιβλεπόμενη μάθηση, δηλαδή δίνονται οι x τιμές στις y κατηγορίες, περιμένοντας κάποιο κελί εξόδου να ενεργοποιηθεί. Τα νευρωνικά τύπου AE, είναι ικανά να εκπαιδευτούν χωρίς επίβλεψη. Αυτά τα νευρωνικά μπορούν και γενικεύουν, αναζητώντας κοινά πρότυπα, όταν ο αριθμός των κρυμμένων νευρώνων είναι μικρότερος από τον αριθμό των νευρώνων εισόδου και όταν ο αριθμός των νευρώνων εξόδου, ισούται με τον αριθμό των νευρώνων εισόδου.

- **Radial Basis Network (RBF)**. Είναι ένα δίκτυο 2 επιπέδων, όπου στο πρώτο επίπεδο εφαρμόζεται η ακτινική λειτουργία βάσης ως συνάρτηση ενεργοποίησης, ενώ στο δεύτερο επίπεδο η γραμμική συνάρτηση. Τα χαρακτηριστικά συνδυάζονται με την ακτινική λειτουργία βάσης στο εσωτερικό στρώμα και στη συνέχεια λαμβάνεται υπ' όψιν, η έξοδος αυτών των χαρακτηριστικών, ενώ υπολογίζεται η ίδια έξοδος, στο επόμενο χρονικό βήμα που είναι βασικά μία μνήμη. Στην πραγματικότητα είναι FF, με κύρια διαφορά ότι χρησιμοποιούν την ακτινική λειτουργία βάσης, έναντι της σιγμοειδούς συνάρτησης.

- **Recurrent Neural Networks (RNN)**. Αυτός ο τύπος νευρωνικού δικτύου, χρησιμοποιείται σε προβλήματα που αποφάσεις και αποτελέσματα από προηγούμενες επαναλήψεις, μπορούν να διαδραματίσουν σημαντικό ρόλο σε καινούρια αποτελέσματα. Οι συνδέσεις μεταξύ των κόμβων, δημιουργούν ένα γράφημα το οποίο είναι κατευθυνόμενο κατά μήκος μίας χρονικής αλληλουχίας. Έτσι, το νευρωνικό τύπου RNN, καθίσταται ικανό να παρουσιάζει χρονική δυναμική συμπεριφορά. Τα RNNs, μπορούν να χρησιμοποιούν την εσωτερική τους κατάσταση (μνήμη) για να επεξεργάζονται ακολουθίες εισόδων και ταξινομούνται σε 4 κατηγορίες με βάση το μήκος των δεδομένων εισόδου και εξόδου. Οι κατηγορίες είναι οι 1-1, 1-πολλά, πολλά-1 και πολλά προς πολλά. Είναι ισχυρά εργαλεία και ικανά να εκπαιδεύονται σε διαδοχικά δεδομένα. Αυτός είναι και ο λόγος που χρησιμοποιήθηκαν στο μοντέλο πρόβλεψης μελλοντικών αναφορών.

- **Long Short Term Memory (LSTM)**. Είναι νευρωνικό τύπου Recurrent και χρησιμοποιείται ευρέως σε εφαρμογές όπως είναι η εγγραφή και η αναγνώριση ομιλίας. Διαθέτει κόμβους που έχουν τη δυνατότητα να θυμούνται και εμφανίζονται ανά 2. Αυτοί οι κόμβοι ονομάζονται πύλες και καθορίζουν τον τρόπο με τον οποίο οι πληροφορίες "κρατιούνται" ή "ξεχνιούνται". Είναι επαναλαμβανόμενα και σε κάθε βήμα αποφασίζουν εάν τα δεδομένα θα προωθηθούν ή όχι. Συγκεκριμένα, η πύλη εισόδου αποφασίζει πόσες πληροφορίες από το τελευταίο δείγμα θα διατηρηθούν στη μνήμη, ενώ η πύλη εξόδου, αποφασίζει την ποσότητα των πληροφοριών που θα μεταφερθούν στο επόμενο στρώμα. Γενικότερα, υπάρχουν πολλές αρχιτεκτονικές των LSTM δικτύων.

4.4.9 Η Αρχιτεκτονική του Encoder-Decoder LSTM Δικτύου

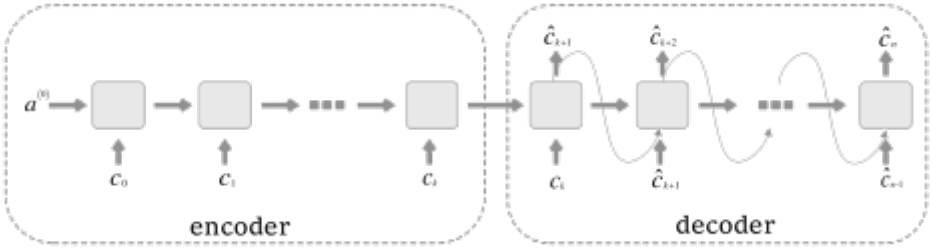
Ο Encoder-Decoder LSTM αναπτύχθηκε για προβλήματα της Ε.Φ.Γ, όπου παρουσίασε υπερσύγχρονη απόδοση, συγκεκριμένα στον τομέα της μετάφρασης κειμένου που ονομάζεται στατιστική μηχανική μετάφραση. Είναι ένα μοντέλο αλληλουχίας που μετατρέπει ακολουθίες από ένα πεδίο σ' ένα άλλο και προσπαθεί να μάθει αποτελεσματικά, το πρότυπο ακολουθίας των αναφορών, ώστε να είναι ικανό να προβλέπει τις μελλοντικές αναφορές. Η πρόβλεψη ακολουθίας, περιλαμβάνει συχνά την πρόβλεψη της επόμενης τιμής σε μία ακολουθία πραγματικής αποτίμησης ή την παραγωγή μίας ετικέτας κλάσης, για μία ακολουθία εισόδου. Αυτό συχνά, πλαισιώνεται ως μία ακολουθία ενός βήματος χρόνου εισόδου, σ' ένα βήμα χρόνου εξόδου (π.χ. ένα προς ένα) ή πολλαπλών βημάτων χρόνου εισόδου, σ' ένα πρόβλημα πρόβλεψης ακολουθίας τύπου βήματος χρόνου (πολλά προς ένα). Υπάρχει και πιο δύσκολος τύπος προβλήματος πρόβλεψης ακολουθίας που λαμβάνει ως είσοδο μία ακολουθία και απαιτεί ως έξοδο μία πρόβλεψη ακολουθίας. Ένα

τέτοιο μοντέλο, ονομάζεται ακολουθία σε μοντέλο αλληλουχίας (seq2seq). Το μήκος των ακολουθιών εισόδου και εξόδου μπορεί να ποικίλει και είναι από τους παράγοντες που μπορούν να δημιουργήσουν προβλήματα. Δεδομένου ότι υπάρχουν πολλαπλά βήματα χρόνου εισόδου και πολλαπλά βήματα χρόνου εξόδου, αυτή η μορφή προβλήματος αναφέρεται ως πρόβλημα πρόβλεψης ακολουθίας σε μοντέλο αλληλουχίας πολλών προς πολλών τύπων.

Η αρχιτεκτονική ενός Encoder-Decoder LSTM δικτύου, αποτελείται από 2 νευρωνικά δίκτυα, τον κωδικοποιητή (Encoder) που εξάγει ένα διάνυσμα πλαισίου (κωδικοποίηση) της ακολουθίας εισόδου, το οποίο στη συνέχεια διαβιβάζεται στον αποκωδικοποιητή (Decoder) που αποτελεί το δεύτερο νευρωνικό δίκτυο, για να αποκωδικοποιήσει και να προβλέψει τις ετικέτες (στόχους). Το LSTM νευρωνικό δίκτυο, απαιτεί τρισδιάστατα δεδομένα, με τον αριθμό των δειγμάτων, των βημάτων χρόνου και των χαρακτηριστικών, να αποτελούν τις 3 διαστάσεις. Ο αριθμός των δειγμάτων είναι ίσος με 3.000, ο αριθμός των χαρακτηριστικών ίσος με 1 και τα βήματα χρόνου ίσα με 3. Η έξοδος, πρέπει επίσης να διαμορφώνεται με αυτόν τον τρόπο όταν χρησιμοποιείται το μοντέλο Encoder-Decoder. Το μήκος της ακολουθίας εισόδου είναι $k+1$ $(c_0, ..., c_k)$, όπου το k ισούται με 12 και το μήκος της ακολουθίας εξόδου είναι $n-k$ $(c_{k+1}, ..., c_n)$, όπου το n ισούται με 15, άρα είναι μεγέθους 3. Τα c_0 έως τα c_k τα λαμβάνει ο κωδικοποιητής ενώ τα c_{k+1} έως τα c_n, είναι οι έξοδοι του τελευταίου και τα παίρνει ο αποκωδικοποιητής για να προβλέψει το c_{k+1}, το οποίο θα χρειαστεί για να προβλεφθεί το c_{k+2} κ.ο.κ. Ένα νευρωνικό δίκτυο, προπονείται αρχικά στη φάση της εκπαίδευσης και στη συνέχεια χρησιμοποιείται για την πρόβλεψη, στη φάση δειγματοληψίας. Στο Σχ. 4.21, αναπαρίσταται η σχηματική περιγραφή της δειγματοληψίας, η οποία έπεται της διαδικασίας της εκπαίδευσης. Σκοπός της τελευταίας είναι η εκπαίδευση του συνόλου δεδομένων εκπαίδευσης, το οποίο χρησιμοποιείται για να μάθει να βελτιστοποιεί τις παραμέτρους του νευρωνικού δικτύου.

Αναλυτικότερα, ο κωδικοποιητής LSTM, λαμβάνει ως είσοδο την ακολουθία της χρονοσειράς μεγέθους 12 και δημιουργεί την αντίστοιχη κωδικοποίηση. Αυτή η κωδικοποίηση αναφέρεται στη μετατροπή των δεδομένων σ' άλλες χωρικές διαστάσεις. Είναι ένα διάνυσμα που αποτελείται απ' όλες τις κρυφές καταστάσεις όλων των νευρώνων του LSTM δικτύου. Στη συνέχεια, μεταφέρεται στον αποκωδικοποιητή LSTM, ως αρχικές καταστάσεις, μαζί με άλλες εισόδους που λαμβάνει ο αποκωδικοποιητής, για να παράγει τις προβλέψεις. Ο αποκωδικοποιητής μετατρέπει ξανά τα δεδομένα στις αρχικές διαστάσεις που είναι τα αποτελέσματα που λαμβάνει ο χρήστης. Οι τελικές προβλέψεις είναι ουσιαστικά οι έξοδοι του αποκωδικοποιητή. Οι ετικέτες που λαμβάνει ως επιθυμητές εξόδους ο αποκωδικοποιητής LSTM, είναι και οι τελευταίοι 3 αριθμοί της κάθε χρονοσειράς που αποτελείται συνολικά από 15 αριθμούς, οι οποίοι συμβολίζουν τον αριθμό αναφορών που έλαβε ένα έγγραφο σε διάρκεια 15 χρόνων.

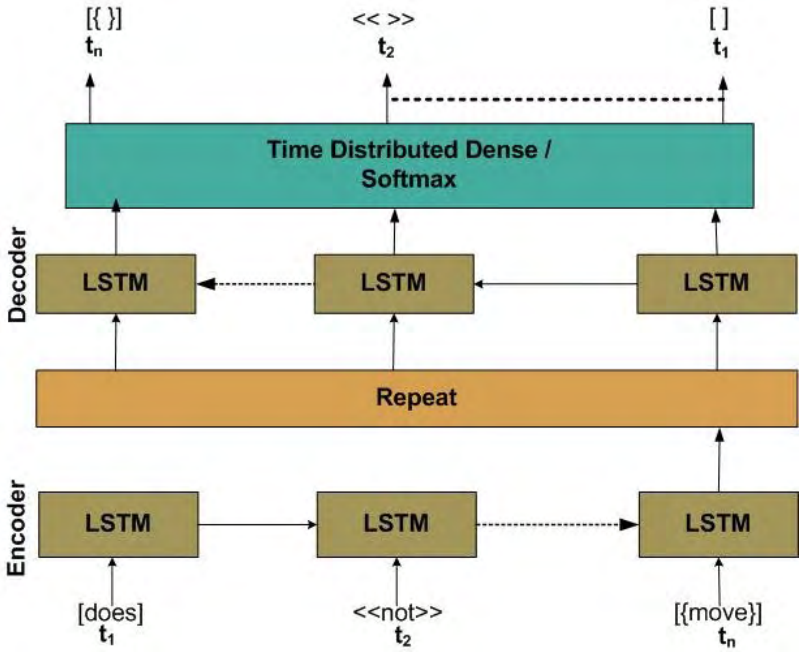
Ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούνται από 100 κρυμμένα στρώματα έκαστος. Αμφότεροι χρησιμοποιούν ως συνάρτηση ενεργοποίησης την ReLU, η οποία εφαρμόζεται σ' όλα τα στρώματα του νευρωνικού δικτύου. Όταν υπάρχουν πολλά στρώματα, οι έξοδοι του πρώτου στρώματος προωθούνται στα επόμενα στρώματα. Έτσι, έχουμε τη δημιουργία ενός LSTM Encoder-Decoder δικτύου, με υψηλά επίπεδα αναπαραστάσεων των δεδομένων της κάθε ακολουθίας εισόδου. Ο κωδικοποιητής LSTM διαβάζει και κωδικοποιεί τις ακολουθίες εισόδου των 3 χρονικών βημάτων. Η κωδικοποιημένη ακολουθία θα επαναληφθεί 3 φορές από το μοντέλο για τα 3 χρονικά βήματα εξόδου που απαιτούνται, χρησιμοποιώντας ένα στρώμα RepeatVector. Αυτά τρο-



Σχήμα 4.21: Διαδικασία δειγματοληψίας για την πρόβλεψη μελλοντικών αναφορών (Πηγή: [4])

φοδοτούνται σ' ένα στρώμα LSTM αποκωδικοποιητή, πριν από τη χρήση ενός στρώματος πυκνής εξόδου που είναι "τυλιγμένο" σ' ένα κατανεμημένου χρόνου στρώμα που θα παράγει μία έξοδο, για κάθε βήμα στην ακολουθία εξόδου. Το κατανεμημένου χρόνου στρώμα, καθιστά τους κόμβους του πυκνού στρώματος πανομοιότυπους και χρησιμοποιείται για τη διατήρηση μίας σχέσης μεταξύ της εισόδου και της εξόδου (ίδια βάρη και μεροληψίες). Η χρήση του είναι συνηθισμένη σε RNN για ταξινόμηση ακολουθιών και με αυτόν τον τρόπο, δίνεται το αποτέλεσμα της ενεργοποίησης της εισόδου, στο επόμενο χρονικό βήμα, κάνοντας τη διαδικασία να μοιάζει με επαναληπτικό βρόχο. Αυτό αποκαλείται κατανεμημένος χρόνος και έχουμε τη δυνατότητα να έχουμε ένα πλήρως συνδεδεμένο πυκνό στρώμα, σε κάθε βήμα χρόνου και να παίρνουμε την έξοδο σε χωριστά χρονικά βήματα. Μετά την εκπαίδευση του νευρωνικού δικτύου χρησιμοποιώντας τα δεδομένα εκπαίδευσης, μπορεί να χρησιμοποιηθεί για να κάνει τις επιθυμητές προβλέψεις.

Στο Σχ. 4.22, έχουμε ένα παράδειγμα ενός Encoder-Decoder LSTM δικτύου, όπου ο κωδικοποιητής αποτελείται από 3 κρυμμένα στρώματα LSTM, από ισάριθμα κρυμμένα στρώματα ο αποκωδικοποιητής και από ένα κατανεμημένου χρόνου στρώμα πριν την τελική έξοδο. Παρατηρούμε ότι η είσοδος του κωδικοποιητή δεν είναι ακολουθίες χρόνου αλλά ακολουθίες λέξεων. Η είσοδος για όλα τα βήματα στον αποκωδικοποιητή, είναι η ίδια και είναι μία κωδικοποιημένη έκδοση όλων των κρυφών καταστάσεων του κωδικοποιητή. Επίσης, μεταξύ του κωδικοποιητή και του αποκωδικοποιητή, παρατηρείται το RepeatVector αλλά και η συνάρτηση Softmax που αποτελεί τη συνάρτηση ενεργοποίησης του δικτύου του παραδείγματος. Τέλος, βλέπουμε ότι το μήκος της ακολουθίας εισόδου είναι ισάριθμο με το μήκος της ακολουθίας εξόδου και ισούται με n.



Σχήμα 4.22: Παράδειγμα ενός Encoder-Decoder LSTM δικτύου (Πηγή: [24])

Κεφάλαιο 5

Ανάπτυξη Μοντέλων και Διεξαγωγή Πειραμάτων

Στο κεφάλαιο αυτό γίνεται η παρουσίαση των αποτελεσμάτων των αλγορίθμων που εφαρμόστηκαν και η σύγκριση απόδοσης αυτών, τροποποιώντας παραμέτρους των αναπτυχθέντων μοντέλων. Επίσης συμπεριλαμβάνονται τα λογισμικά που χρησιμοποιήθηκαν για τη μελέτη και επίλυση των προβλημάτων.

5.1 Μοντέλο LDA

5.1.1 Λεπτομέρειες Εκπαίδευσης του Αλγορίθμου LDA

Το γεγονός ότι τα περισσότερα δεδομένα είναι σε αδόμητη μορφή, χωρίς κάποια συγκεκριμένη οργάνωση ή μορφή, καθιστά την ανάγκη για μία σωστή και αποτελεσματική προεπεξεργασία αυτών, με απώτερο σκοπό τη μεγαλύτερη απόδοση του μοντέλου. Η συντακτική και η σημασιολογική ανάλυση είναι από τις κύριες τεχνικές της Ε.Φ.Γ. Η συντακτική ανάλυση αναφέρεται στη διάταξη των λέξεων σε μία πρόταση, με απώτερο σκοπό την απόκτηση γραμματικής έννοιας, ενώ η σημασιολογική ανάλυση, αναφέρεται στο νόημα που χαρακτηρίζει ένα κείμενο. Κύριες τεχνικές της συντακτικής ανάλυσης είναι η λημματοποίηση, η οποία μειώνει τις διαστάσεις των μορφών των λέξεων και τις επαναφέρει στη ριζική τους μορφή για πιο εύκολη ανάλυση, η αφαίρεση σημείων στίξης κ.λπ. Από την άλλη, στο πεδίο της σημασιολογικής ανάλυσης που είναι μία από τις πιο δύσκολες πτυχές της Ε.Φ.Γ, συναντώνται τεχνικές που αφορούν την απόδοση νοήματος σε μία λέξη που βασίζεται στα συμφραζόμενα, την εξαγωγή σημασιολογικών προθέσεων και τη μετατροπή τους στην ανθρώπινη γλώσσα κ.λπ.

Για την επίδειξη της μοντελοποίησης των θεμάτων από τις περιλήψεις των εγγράφων, αξιοποιήθηκαν 15.000 έγγραφα από το αρχικό σύνολο των επιστημονομετρικών δεδομένων. Τα κυριότερα βήματα για την προεπεξεργασία κειμένου, προτού αναλυθεί εκτενέστερα ολόκληρη η διαδικασία, είναι τα εξής. Η αφαίρεση σημείων στίξης, η λημματοποίηση, ο καθορισμός επιτρεπόμενων μερών του λόγου, το φιλτράρισμα λέξεων και η μετατροπή των κεφαλαίων σε πεζά. Αναλυτικότερα, αφαιρέθηκαν από το ακατέργαστο σύνολο κειμένων (τις περιλήψεις), τα σημεία στίξης όπως τα

κόμματα, τα θαυμαστικά καθώς και οι κενοί χαρακτήρες. Στη συνέχεια, αξιοποιώντας τεχνικές της συντακτικής ανάλυσης, όλες οι προτάσεις μετατράπηκαν σε ξεχωριστές μονάδες (λέξεις) και ως επιτρεπόμενα μέρη του λόγου ορίστηκαν τα ουσιαστικά, τα ρήματα, τα επιρρήματα και τα επίθετα. Επιπρόσθετα, αξιοποιήθηκε το φιλτράρισμα διακοπής λέξεων, όπου αφαιρούνται όλοι οι όροι που εμφανίζονται σε μία προκαθορισμένη λίστα, αποτελούμενη από συγκεκριμένους όρους που δεν συνεισφέρουν χρήσιμες πληροφορίες σχετικά με την εξαγωγή των θεμάτων. Επίσης, όλα τα κεφαλαία γράματα μετατράπηκαν σε πεζά και αφαιρέθηκαν λέξεις που είχαν μήκος μικρότερο του 3, αλλά και αυτές που εμφανίζονται σε λιγότερα από 100 ή περισσότερα από 1000 έγραφα. Εν συνεχεία, εφαρμόστηκε η λημματοποίηση στα δεδομένα μας. Τέλος, πριν παραλάβει τα δεδομένα ο αλγόριθμος της μοντελοποίησης των θεμάτων, εφαρμόστηκε το μοντέλο Bag-of-Words, το οποίο επικεντρώνεται στην εμφάνιση των πιο σχνών λέξεων μέσα σ' ένα έγγραφο. Μετατρέπει όλους τους όρους σ' έναν πίνακα που περιέχει όλες τις μοναδικές λέξεις των περιλήψεων και τις φορές που εμφανίζονται σε κάθε έγγραφο ξεχωριστά. Ουσιαστικά κάθε έγγραφο, αντιπροσωπεύεται από ένα διάνυσμα όρο, με μία καταχώρηση που δηλώνει τον αριθμό των φορών που εμφανίζεται ο όρος στο έγγραφο. Ο πίνακας αυτός ονομάζεται πίνακας εγγράφου-όρου, αποτελείται από τις συχνότητες των όρων, είναι δισδιάστατος, αραιός και οι γραμμές του αντιστοιχούν στα έγγραφα, ενώ οι στήλες του στους αντίστοιχους όρους (λέξεις). Ένα τέτοιο παράδειγμα πίνακα, απεικονίζεται στον Πίνακα 5.1. Πλέον, μπορούμε να έχουμε πρόσβαση σε όλους τους όρους, αφού ο πίνακας χαρτογραφεί κάθε μοναδικό όρο στην αντίστοιχη στήλη. Οι περιλήψεις των 15.000 εγγράφων που χρησιμοποιήθηκαν για το συγκεκριμένο πείραμα, αποτελούνταν από 1.856.159 λέξεις στο σύνολο, ενώ από αυτές οι 565.436 ήταν μοναδικές. Μετά από αυτή την επεξεργασία που προαναφέραμε, έχουμε έναν πίνακα με διαστάσεις 15.000 x 1.316, δηλαδή ο πίνακας διαθέτει 15.000 γραμμές (έγγραφα) και 1.316 στήλες (λέξεις-όρους).

	term 1	term 2	term 3	term 4	term 5
doc 1	0	0	1	0	0
doc 2	2	0	1	1	0
doc 3	0	0	1	1	0
doc 4	0	0	1	1	1

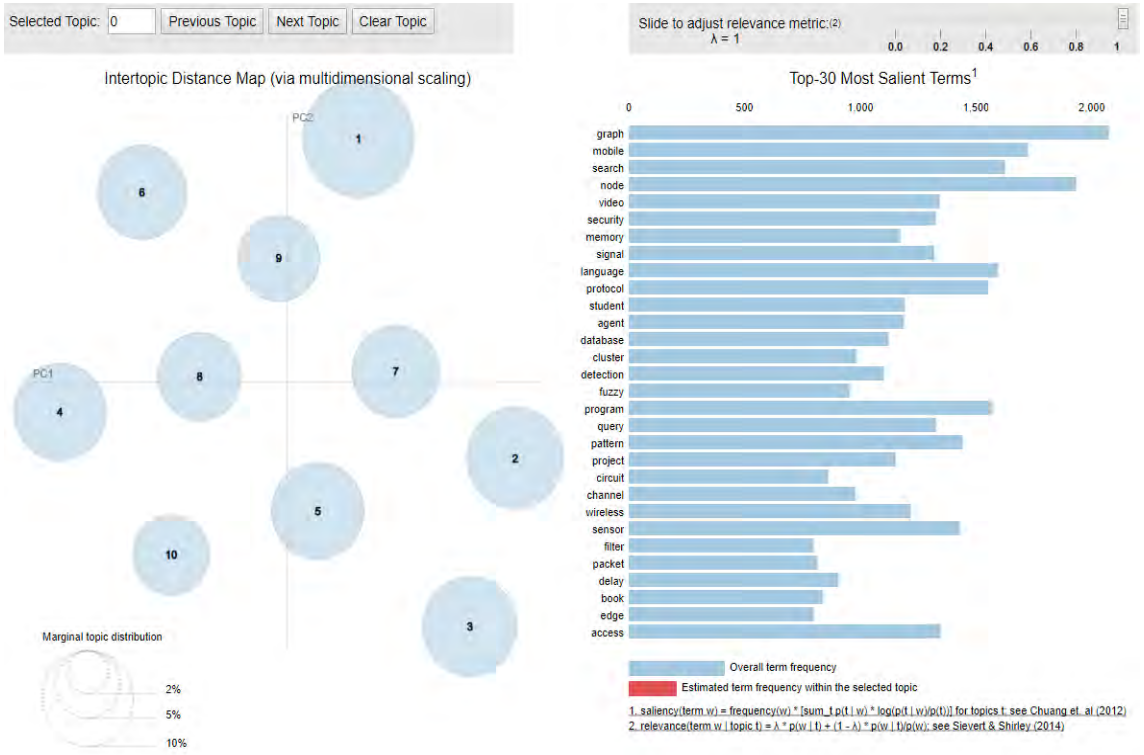
Πίνακας 5.1: Παράδειγμα πίνακα εγγράφου-όρου

Όσον αφορά τις παραμέτρους του LDA αλγορίθμου, το μέγεθος παρτίδας έχει οριστεί σε 1024. Αυτό το μέγεθος ουσιαστικά ορίζει τον αριθμό των δειγμάτων που θα διαδοθούν μέσω του δικτύου. Για παράδειγμα, ας πούμε ότι έχουμε 2000 δείγματα εκπαίδευσης και θέλουμε να ρυθμίσουμε το μέγεθος της παρτίδας να ισούται με 100. Ο αλγόριθμος παίρνει τα πρώτα 100 δείγματα από το σύνολο δεδομένων εκπαίδευσης και εκπαιδεύει το μοντέλο. Στη συνέχεια, παίρνει τα δεύτερα 100 δείγματα και συνεχίζει η εκπαίδευση του μοντέλου με τα αντίστοιχα δείγματα. Μπορούμε να συνεχίσουμε να κάνουμε αυτή τη διαδικασία μέχρι να διαδοθούν όλα τα δείγματα στο μοντέλο. Ο αριθμός θεμάτων που επιλέξαμε να παράγει, ισούται με 10. Επίσης ο ρυθμός φθοράς (πτώσης) της μά-

θησης του μοντέλου ισούται με 0.7. Πρόκειται για μία παράμετρο που ελέγχει τον ρυθμό μάθησης στην ηλεκτρονική μέθοδο μάθησης που επιλέχθηκε για το μοντέλο. Η τιμή πρέπει να οριστεί μεταξύ (0,5-1) για να εξασφαλιστεί ασυμπτωτική σύγκλιση. Ο τύπος που συνδέει τον παράγοντας της πτώσης με τον ρυθμό μάθησης του μοντέλου είναι: $lrate = initiallrate*(1/(1+decay*iteration))$. Όσο μεγαλύτερος είναι αυτός ο παράγοντας, τόσο πιο πολύ μειώνεται ο ρυθμός μάθησης εντός των ίδιων επαναλήψεων. Η επιλογή της ηλεκτρονικής μάθησης έναντι της μάθησης ανά παρτίδα, είναι η καλύτερη όταν πρόκειται για μεγάλο σύνολο δεδομένων. Τέλος, ο μέγιστος αριθμός επαναλήψεων του αλγορίθμου, έχει οριστεί σε 20, ενώ η αντιστάθμιση μάθησης, η οποία είναι μία (θετική) παράμετρος μεγαλύτερη από 1 και μειώνει τις αρχικές επαναλήψεις στην ηλεκτρονική μάθηση, έχει οριστεί σε 10.

5.1.2 Σχηματική Απεικόνιση των Θεμάτων και των Λέξεων-Κλειδιών του LDA

Μία αποτελεσματική μοντελοποίηση θεμάτων, χαρακτηρίζεται από μη επικαλυπτόμενους μεγάλους κύκλους που θα συμβολίζει ο καθένας και από ένα θέμα αντίστοιχα. Στο Σχ. 5.1, παρουσιάζονται τα αποτελέσματα της μοντελοποίησης θεμάτων από 15.000 δείγματα περιλήψεων, του αρχικού συνόλου δεδομένων που αξιοποιήθηκαν για το πείραμα αυτό. Στα αριστερά, σε σχήμα ενός κύκλου, απεικονίζεται το κάθε θέμα. Όσο μεγαλύτερη είναι η επιφάνεια του κύκλου, τόσο επικρατέστερο είναι και το αντίστοιχο θέμα. Όσο πιο διάσπαρτοι είναι οι κύκλοι σ' ολόκληρο το γράφημα, αντί να είναι συγκεντρωμένοι σ' ένα τεταρτημόριο, τόσο το καλύτερο. Ένα μη αποτελεσματικό μοντέλο, με παραπάνω εξαγόμενα θέματα απ' αυτά που εκπροσωπούσε, θα είχε ως αποτελέσμα τη δημιουργία μικρών και επικαλυπτόμενων κύκλων, συγκεντρωμένων σε μία περιοχή του γραφήματος. Στη δεξιά περιοχή του γραφήματος, φαίνονται οι λέξεις-κλειδιά και η συνολική συχνότητα αυτών στα θέματα που εντοπίστηκαν. Το γράφημα, σε προγραμματιστικό περιβάλλον είναι διαδραστικό, δηλαδή αν πατήσουμε πάνω σ' ένα θέμα, στα δεξιά μας θα εμφανίσει τη συχνότητα των λέξεων του αντίστοιχου θέματος. Μετά από το πείραμα αξιολόγησης του μοντέλου [5.1.7] για διαφορετικές τιμές θεμάτων, ο αριθμός θεμάτων που δημιουργεί σαφή και μη επικαλυπτόμενα θέματα, ήταν 10. Η αρίθμηση των θεμάτων στην παρακάτω απεικόνιση 5.1, δεν σχετίζεται με την αρίθμηση αυτών στα υπόλοιπα σχήματα. Το μοντέλο LDA εξάγει τις πιο σημαντικές λέξεις-κλειδιά που αντιστοιχούν στο κάθε θέμα ξεχωριστά. Στο Σχ. 5.2, φαίνονται οι 7 λέξεις με το μεγαλύτερο βάρος, δηλαδή οι λέξεις που έχουν τη μεγαλύτερη επιρροή στον σχηματισμό του αντίστοιχου θέματος. Για παράδειγμα, αν δούμε τους πιο σημαντικούς όρους του θέματος 7 (Topic 6), συμπεραίνουμε ότι το θέμα λογικά αναφέρεται στην αρχιτεκτονική ενός επεξεργαστή γραφημάτων.



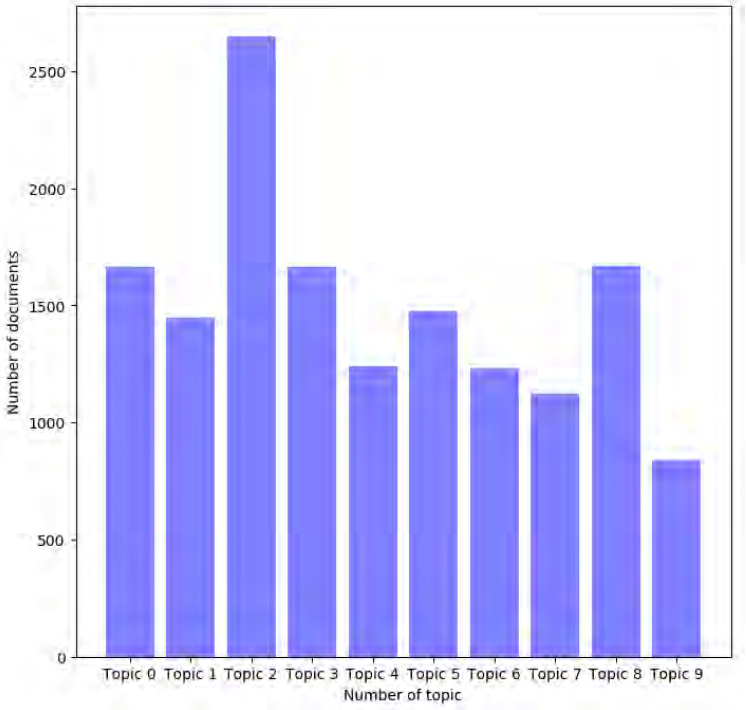
Σχήμα 5.1: Απεικόνιση θεμάτων και λέξεων-κλειδιών του LDA

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Topic 0	human	interaction	social	activity	people	market	world
Topic 1	mobile	security	protocol	access	sensor	grid	policy
Topic 2	equation	matrix	approximation	numerical	variable	distribution	error
Topic 3	video	object	surface	motion	region	frame	resolution
Topic 4	pattern	database	query	game	recognition	document	text
Topic 5	student	agent	project	book	group	business	course
Topic 6	graph	memory	circuit	edge	processor	parallel	vertex
Topic 7	node	channel	delay	packet	traffic	path	wireless
Topic 8	language	program	programming	component	robot	machine	logic
Topic 9	search	signal	detection	cluster	fuzzy	filter	noise

Σχήμα 5.2: Οι λέξεις-κλειδιά για κάθε θέμα

5.1.3 Κατανομή των Θεμάτων Μεταξύ των Εγγράφων

Στο Σχ. 5.3, βλέπουμε το συνολικό αριθμό εγγράφων που διαθέτει το κάθε θέμα. Παρατηρούμε ότι το θέμα 3 (Topic 2), διαθέτει το υψηλότερο σύνολο εγγράφων που ξεπερνάει τα 2500, ενώ τα υπόλοιπα θέματα, απαρτίζονται από 700 μέχρι 1700 περίπου έγγραφα.



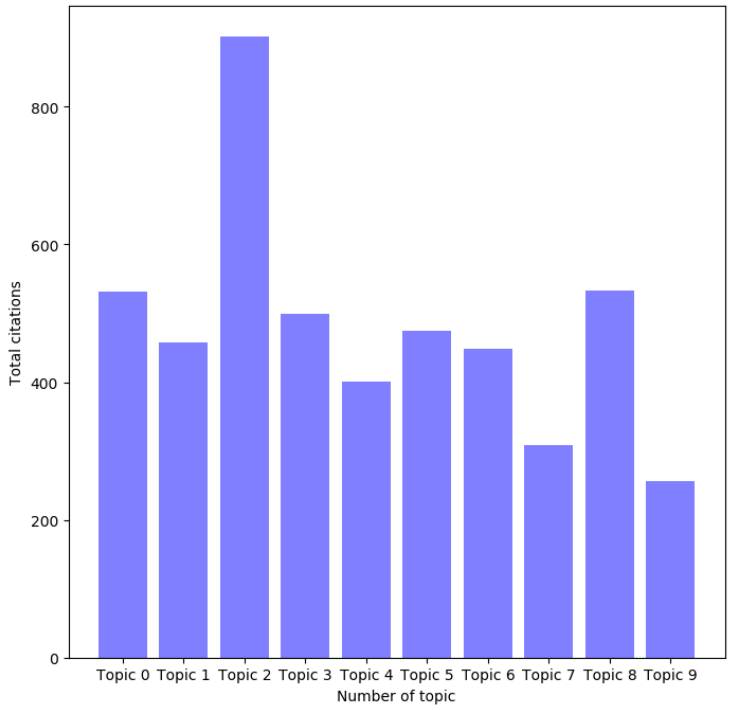
Σχήμα 5.3: Αριθμός εγγράφων που διαθέτει το κάθε θέμα

5.1.4 Συνολικές Αναφορές των Εγγράφων του κάθε Θέματος

Στο Σχ. 5.4, βλέπουμε τις συνολικές αναφορές που έλαβαν τα έγγραφα που απαρτίζουν το κάθε θέμα. Συγκεκριμένα, το θέμα 3 (Topic 2) έλαβε τις περισσότερες αναφορές την περίοδο 1900-2009, με το συνολικό αριθμό των αναφορών να ξεπερνάει το 800. Τα υπόλοιπα θέματα κυμαίνονται περίπου στις ίδιες τιμές αναφορών την ίδια χρονική περίοδο, με τις αναφορές να ξεκινούν από 250 και να φτάνουν τις 530 περίπου.

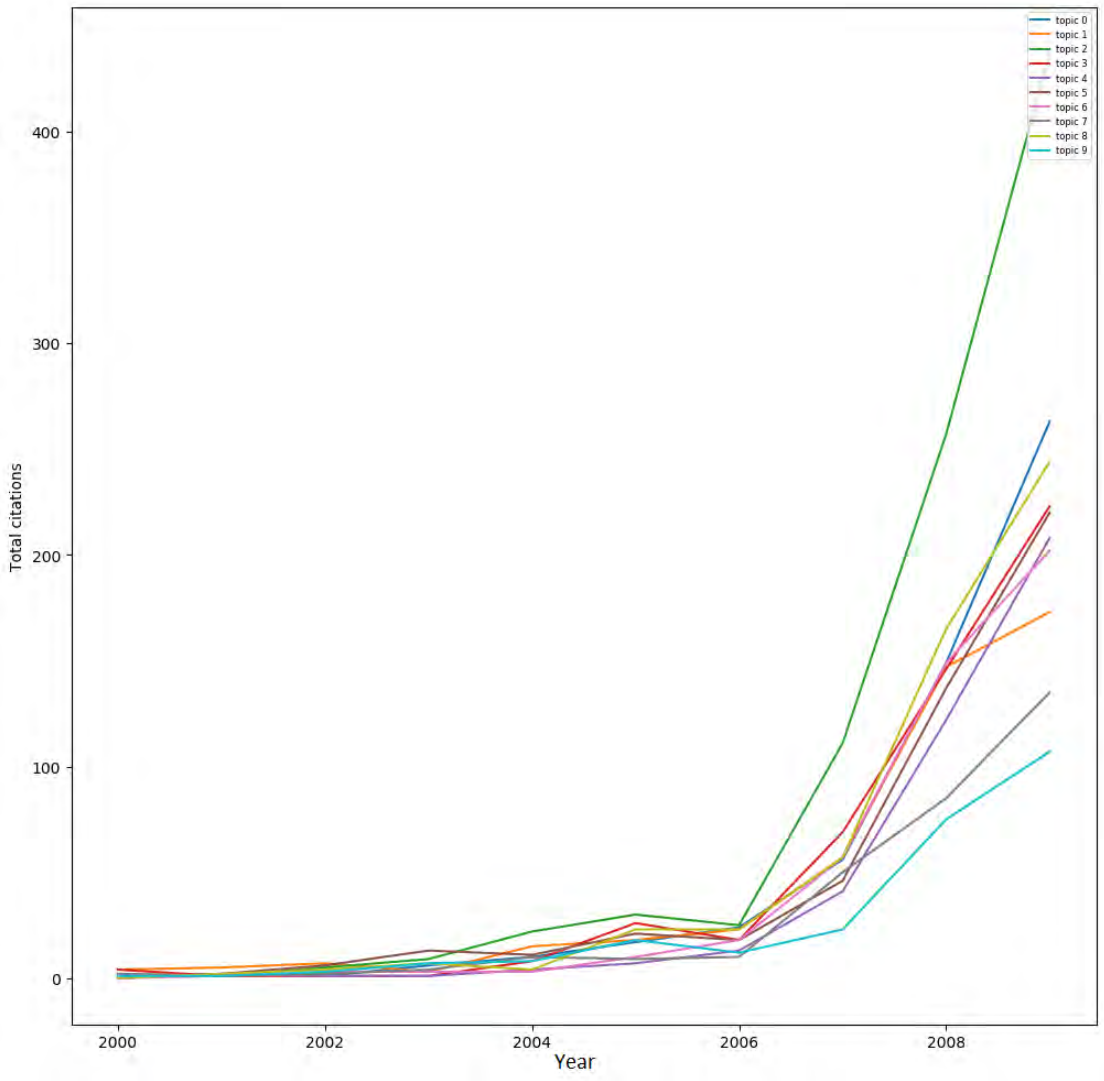
Παράδειγμα Κατανομής των Συνολικών Αναφορών των Θεμάτων τη Περίοδο 2000-2009

Μία επιπλέον χρήσιμη πληροφορία που μπορούμε να εξάγουμε από τα επιστημονομετρικά μας δεδομένα, είναι η απήχηση των αναφερόμενων ή μη, εγγράφων που αποτελούν το κάθε θέμα, στο πέρασμα του χρόνου. Με αυτόν τον τρόπο, μπορούμε να δούμε ποιο ή ποια από τα θέματα ήταν τα πιο αναφερόμενα και απασχόλησαν την επιστημονική κοινότητα. Στο Σχ. 5.5, παρατηρούμε ότι όλα τα θέματα από το 2000 μέχρι και λίγο πριν το 2004, είχαν λάβει αναφορές που δεν ξεπερνούσαν τις 10 με 15. Την περίοδο 2004 έως 2006, διαπιστώνουμε μία “αδύναμη” ανοδική πορεία των θεμάτων, ενώ από το 2006 έως και το 2009 υπήρξε μία σημαντική αύξηση στο συνολικό αριθμό



Σχήμα 5.4: Συνολικός αριθμός αναφορών ανά θέμα την περίοδο 2000-2009

των παραπομπών των εγγράφων τους. Συγκεκριμένα, ραγδαία άνοδο είχε το θέμα 3 (Topic 2) που εκπροσωπείται από την πράσινη γραμμή, με τις αναφορές να ξεπερνούν τις 400 το 2010, ενώ τη μικρότερη ανοδική πορεία φαίνεται να είχε το θέμα 8 (Topic 7) και το θέμα 10 (Topic 9) που εκπροσωπούνται από τη γκρι και τη γαλάζια γραμμή αντίστοιχα. Τα υπόλοιπα θέματα "εμφάνισαν" παραπλήσια άνοδο, στις παραπομπές που έλαβαν τα έγγραφα που τα απαρτίζουν.



Σχήμα 5.5: Παράδειγμα κατανομής των συνολικών αναφορών που έλαβαν τα εξαγόμενα θέματα την περίοδο 2000-2009

5.1.5 Δημιουργία Πίνακα Πιθανοτήτων Εγγράφου-Θέματος

Μετά τη μοντελοποίηση θεμάτων που προηγήθηκε, κάθε έγγραφο ανήκει σ' ένα θέμα. Θα δημιουργήσουμε έναν πίνακα εγγράφου-θέματος που θα δείχνει για κάθε έγγραφο, ποια ήταν η πιθανοτική συμβολή του στο κάθε θέμα αντίστοιχα. Με μπλε χρώμα, συμβολίζονται όλα τα θέματα που είναι τα πιο σημαντικά για κάθε έγγραφο. Στο Σχ. 5.6, οι γραμμές αντιπροσωπεύουν τα έγγραφα και οι στήλες τα θέματα που δημιουργήθηκαν. Συγκεκριμένα, ο πίνακας είναι ένα δείγμα, αποτελούμενος από τα πρώτα 10 έγγραφα των 15.000 που αξιοποιήθηκαν για τη μοντελοποίηση, με τις διαστάσεις του να είναι 15.000 x 10, όπου 15.000 ο αριθμός των εγγράφων και 10 ο αριθμός των παραγόμενων θεμάτων. Η ποσοτικοποίηση της συμβολής όλων των εγγράφων σε κάθε θέμα, ήταν ουσιαστικά η παραγόμενη έξοδος του αλγορίθμου LDA και αποτέλεσε τη βάση για τη συσταδοποίηση εγγράφων [5.2.1] που "μοιράζονται" παρόμοια θέματα.

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
Doc0	7.45	4.46	0	0	0	0	2.73	5.43	6.1	3.43
Doc1	10.37	10.07	0	3.6	0	0	0	0	14.35	0
Doc2	0	0	0	0	11.72	5.96	0	0	6.62	0
Doc3	0	11.3	0	0	0	0	12.37	21.63	0	0
Doc4	0	34.16	4.22	0	0	0	0	0	0	5.92
Doc5	8.86	0	3.74	0	4.63	6.17	0	0	0	0
Doc6	0	0.71	0	10.98	20.27	0	0	2.06	0	17.48
Doc7	0	0	0	7.07	0	44.13	0	1.91	3.28	0
Doc8	0	0	29.1	0	0	0	0	0	0	0
Doc9	4.01	0	2.37	0	14.09	0	0	0	4.93	0

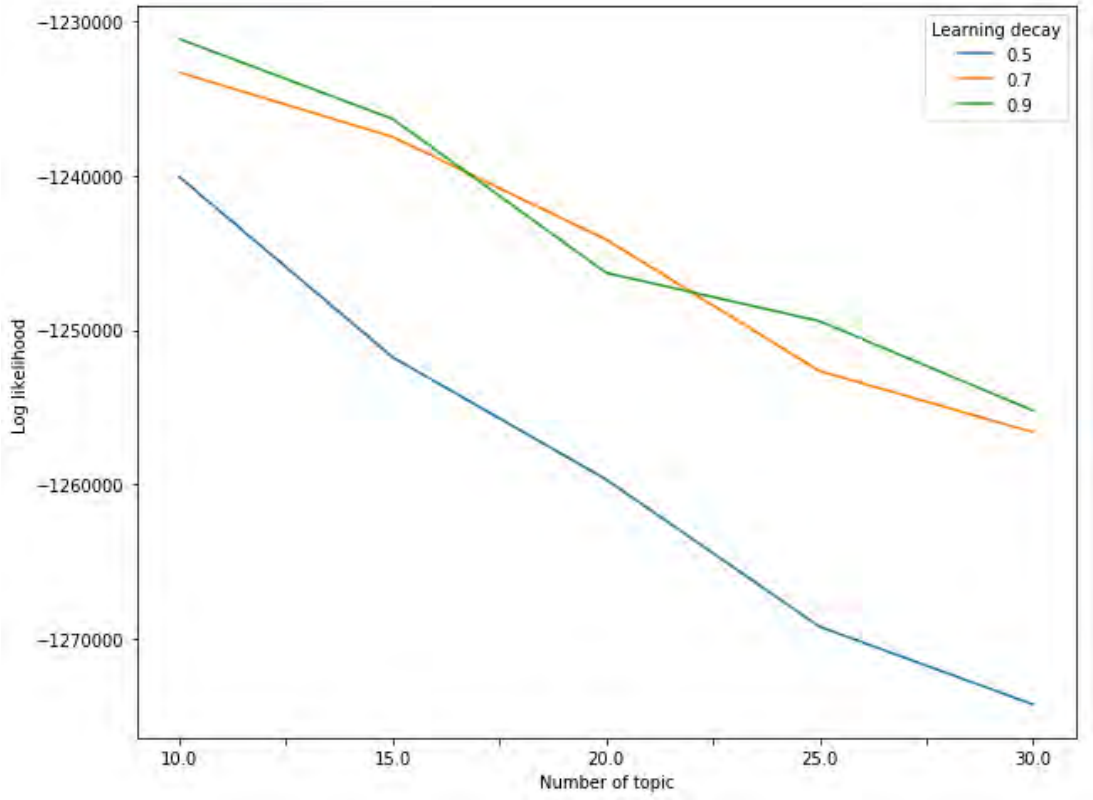
Σχήμα 5.6: Απεικόνιση πιθανοτήτων (%) εγγράφου-θέματος

5.1.6 Μέτρα Αξιολόγησης του Αλγορίθμου LDA

Ο αλγορίθμος LDA αξιολογήθηκε με βάση 2 μετρικά. Το ένα μετρικό αποκαλείται λογαριθμική πιθανότητα (log-likelihood) και είναι ο λογαριθμικός μετασχηματισμός της συνάρτησης πιθανότητας. Δεδομένων των ανεξάρτητων γεγονότων, η συνολική λογαριθμική πιθανότητα είναι το άθροισμα των λογαριθμικών πιθανοτήτων των μεμονωμένων γεγονότων. Όσο μεγαλύτερες οι τιμές που παράγει, τόσο το καλύτερο για το εκάστοτε μοντέλο. Το μοντέλο μας, με τη δημιουργία 10 θεμάτων, έχει λογαριθμική πιθανότητα ίση με -3.570.545,05. Το άλλο μετρικό που χρησιμοποιήθηκε για την αξιολόγηση, είναι η λεγόμενη σύγχυση (perplexity). Ο τύπος του τελευταίου μετρικού είναι: $perplexity = \exp(-1 * loglikelihood)$ ανά λέξη. Όσο μικρότερες τιμές έχει αυτή η συνάρτηση, τόσο καλύτερη είναι η απόδοση του αλγορίθμου. Στην περίπτωση του μοντέλου μας, η σύγχυση ισούται με 853.42.

5.1.7 Τροποποίηση Παραμέτρων του Αλγορίθμου LDA

Παρακάτω, στο Σχ. 5.7, απεικονίζεται η γραφική αναπαράσταση της Log-likelihood για διαφορετικό αριθμό θεμάτων, τροποποιώντας ταυτόχρονα τον ρυθμό πτώσης του μοντέλου. Το πείραμα πραγματοποιήθηκε με τον εξεταζόμενο αριθμό θεμάτων να είναι 10, 15, 20, 25 και 30 και τον ρυθμό πτώσης να είναι 0.5, 0.7 και 0.9. Όπως αναφέραμε, θέλουμε τη λογαριθμική πιθανότητα να παράγει μεγάλες τιμές. Όπως φαίνεται και στο σχήμα, τη μεγαλύτερη τιμή την είχε η πράσινη καμπύλη που αφορά τον ρυθμό φθοράς που ισούται με 0.9 και σχετίζεται με 10 θέματα. Έτσι λοιπόν, μπορούμε να χρησιμοποιήσουμε τον παραπάνω ρυθμό φθοράς. Ο αριθμός θεμάτων είναι 10, όπως επιλέχθηκε και στο δικό μας μοντέλο. Συνεπώς, εκτός από τον ρυθμό πτώσης που θα μπορούσε να αλλάξει από 0.7 σε 0.9, αν και δεν υπάρχει σημαντική διαφορά, το βέλτιστο μοντέλο έχει τις παραμέτρους που χρησιμοποιήσαμε στο τελικό μας μοντέλο. Η σύγχυση στο βέλτιστο μοντέλο ισούται με 754.98 και η λογαριθμική πιθανότητα με -1.231.114,75.



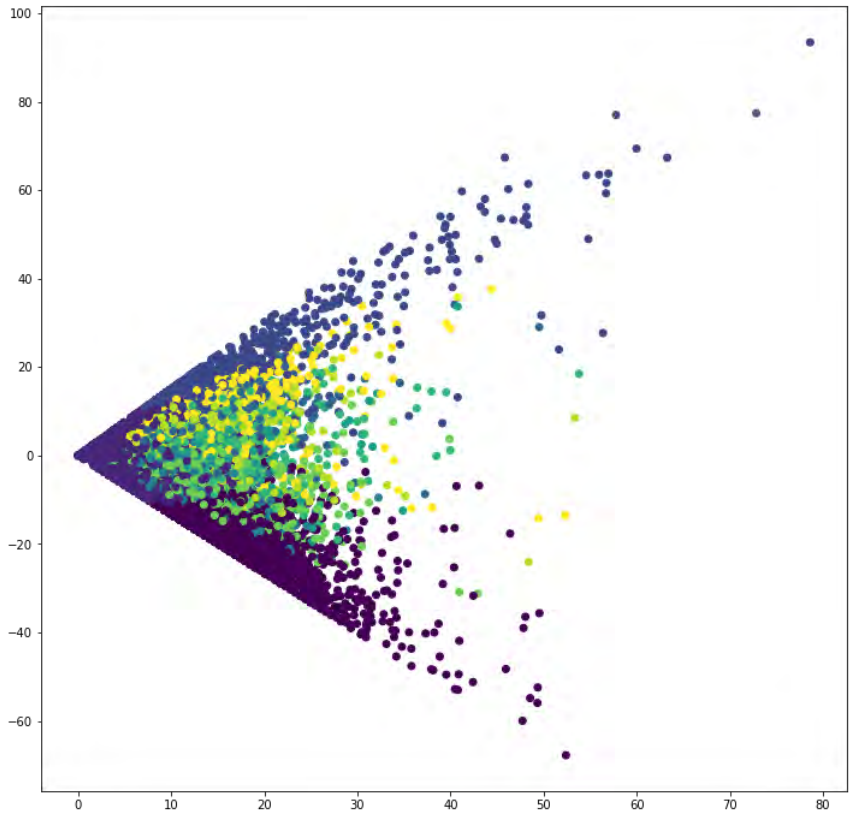
Σχήμα 5.7: Απεικόνιση βέλτιστου μοντέλου LDA

5.2 Μοντέλο K-Means

5.2.1 Συσταδοποίηση Εγγράφων που Μοιράζονται Παρόμοια Θέματα

Ο αλγόριθμος K-Means εφαρμόζεται στην έξοδο του LDA που είναι ο πίνακας πιθανοτήτων [5.6] εγγράφου-θέματος, διαστάσεων 15.000 x 10. Για την απεικόνιση της συσταδοποίησης των

εγγράφων που "μοιράζονται" παρόμοια θέματα, χρειαζόμαστε τις συντεταγμένες x και y. Για να επιτευχθεί η γραμμική μείωση των διαστάσεων όλων των δεδομένων, εφαρμόστηκε η μέθοδος της Αποσύνθεσης Μοναδικής Τιμής στον προαναφερόμενο πίνακα, με επιθυμητή διάσταση των δεδομένων εξόδου να είναι 2. Η προαναφερόμενη μέθοδος, διασφαλίζει ότι αυτές οι 2 στήλες θα καταγράψουν τη μέγιστη δυνατή ποσότητα πληροφοριών από την έξοδο του LDA, στα 2 πρώτα στοιχεία. Στο Σχ. 5.8, είναι το αποτελέσμα της συσταδοποίησης των εγγράφων, όπου απεικονίζεται ο διαχωρισμός των θεματικών συστάδων. Κατά μήκος των 2 στοιχείων της διαδικασίας της αποσύνθεσης είναι τα έγγραφα, όπου το χρώμα των σημείων αντιπροσωπεύει τον αριθμό της συστάδας ή αλλιώς στη συγκεκριμένη περίπτωση, τον αριθμό του θέματος. Το πλήθος των συστάδων είναι 10 και αντιστοιχεί στον αριθμό των θεμάτων που προέκυψαν από τη διαδικασία της μοντελοποίηση θεμάτων. Ο άξονας x αντιπροσωπεύει το 1^o στοιχείο και ο άξονας y το 2^o στοιχείο της μεθόδου της Αποσύνθεσης Μοναδικής Τιμής.



Σχήμα 5.8: Απεικόνιση συσταδοποίησης εγγράφων που μοιράζονται παρόμοια θέματα

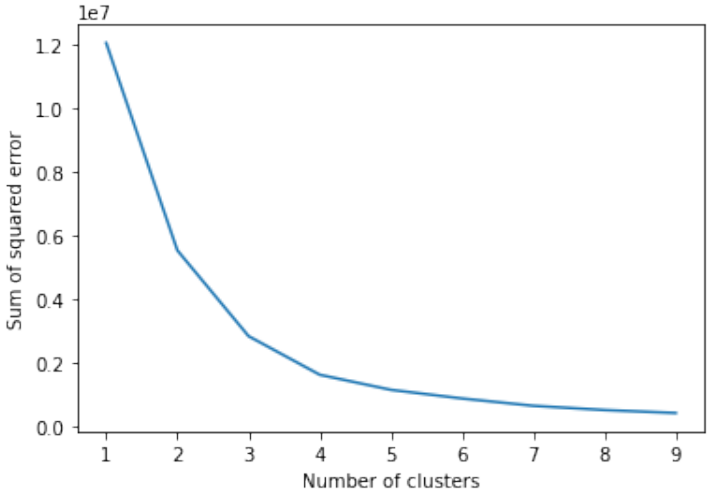
5.2.2 Συσταδοποίηση Εγγράφων για τον Εντοπισμό Σχέσης Αναφορών-Τίτλων

Σε αυτήν την υποενότητα, θα προσπαθήσουμε να πραγματοποιήσουμε συσταδοποίηση εγγράφων, με συντεταγμένες το πλήθος των παραπομπών που έλαβαν την περίοδο 1900-2009 και το μέγεθος των τίτλων τους. Θα χρησιμοποιήσουμε ένα δείγμα από το αρχικό σύνολο δεδομένων μας, το οποίο θα αποτελείται από περίπου 12.000 έγγραφα. Στο Σχ. 5.9, απεικονίζεται το κριτήριο Elbow, το οποίο εφαρμόζεται με απώτερο σκοπό τον εντοπισμό του καταλληλότερου αριθμού συ-

στάδων. Όπως παρατηρούμε, ο "αγκώνας" δημιουργείται μεταξύ αριθμών 2 και 3. Θα μπορούσε να επιλεχθεί και ο αριθμός 2, αλλά επιλέχθηκε ο αριθμός 3. Το εύρος των εξεταζόμενων συστάδων ήταν από 1 ως 10. Ο άξονας x συμβολίζει τον αριθμό των συστάδων, ενώ ο άξονας y, συμβολίζει το άθροισμα των τετραγωνικών σφαλμάτων (SSE), το οποίο αποτελεί ένα μέτρο της απόκλισης μεταξύ των δεδομένων και του μοντέλου εκτίμησης. Ο συμβολισμός $1e7$ είναι μία τυπική επιστημονική έννοια και εδώ υποδεικνύει έναν συνολικό συντελεστή κλίμακας για τον άξονα y. Δηλαδή το 0.8 που υπάρχει στον άξονα y, ουσιαστικά η τιμή του είναι $0.8 * 10^7$. Μικρές τιμές της συνάρτησης, συμβολίζουν στενή εφαρμογή του μοντέλου στα δεδομένα. Η Εξίσωση 5.1, αποτελεί τον μαθηματικό τύπο του παραπάνω μέτρου.

$$SSE = \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{5.1}$$

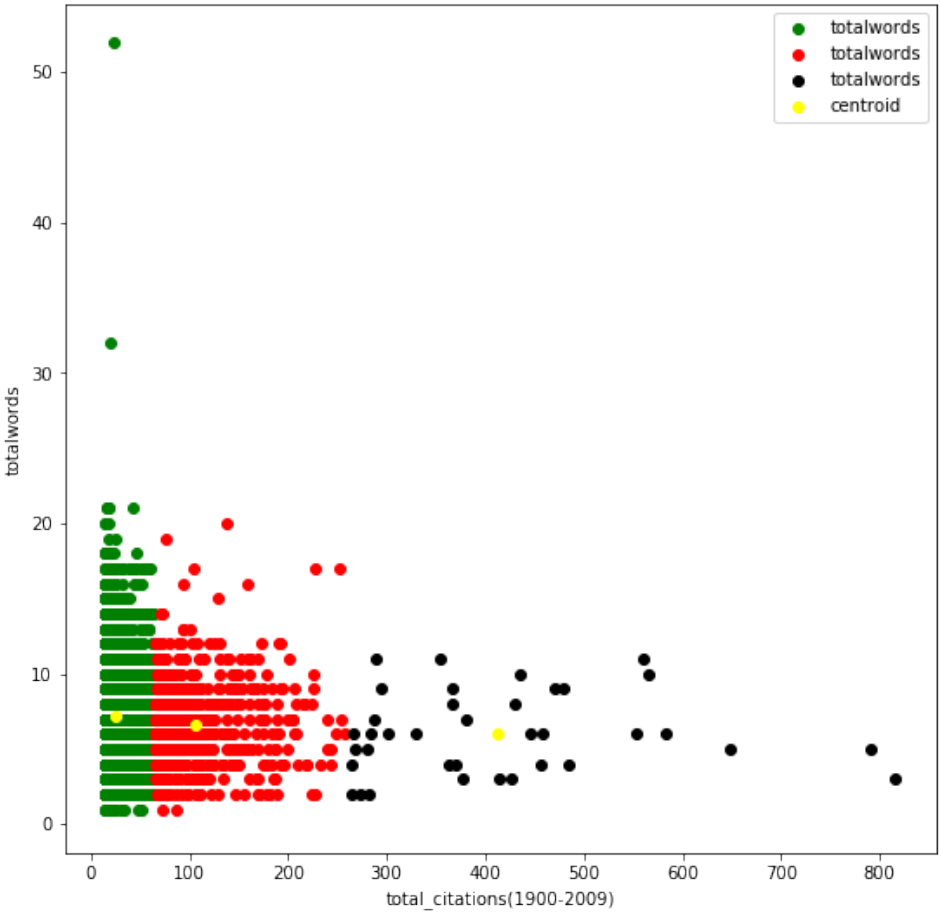
όπου n είναι ο αριθμός των δεδομένων, y_i είναι η i-οστή παρατηρούμενη τιμή των δεδομένων, x_i είναι η i-οστή τιμή της επεξηγηματικής μεταβλητής και $f(x_i)$ η προβλεπόμενη τιμή από την προσαρμογή της x_i στο μοντέλο. Στον Πίνακα 5.2, δίνονται οι συντεταγμένες των 3 σχηματιζόμενων συστάδων και στο Σχ. 5.10, παρουσιάζεται το αποτέλεσμα της συσταδοποίησης των εγγράφων.



Σχήμα 5.9: Εφαρμογή του κριτηρίου Elbow

	x	y
Πρώτη συστάδα	24.98	7.25
Δεύτερη συστάδα	106.07	6.67
Τρίτη συστάδα	413.08	6.08

Πίνακας 5.2: Οι συντεταγμένες των σχηματιζόμενων συστάδων



Σχήμα 5.10: Απεικόνιση συσταδοποίησης εγγράφων για τον εντοπισμό σχέσης αναφορών-τίτλων

Ο άξονας x συμβολίζει τον αριθμό των αναφορών που έλαβε το κάθε έγγραφο την περίοδο 1900-2009, ενώ ο άξονας y συμβολίζει τον αριθμό των λέξεων που διαθέτει ο τίτλος του αντίστοιχου άρθρου. Με κίτρινο χρώμα συμβολίζονται τα κέντρα των συστάδων, ενώ με πράσινο, κόκκινο και μαύρο χρώμα, η πρώτη, η δεύτερη και η τρίτη συστάδα αντίστοιχα. Από τα κέντρα των ομάδων, παρατηρούμε ότι ο μέσος όρος των αναφορών των εγγράφων της πρώτης συστάδας είναι 24.98 με μέσο όρο μέγεθος τίτλου 7.25 λέξεις. Ο μέσος όρος των αναφορών της δεύτερης συστάδας είναι 106.07, με 6.67 μέσο όρο λέξεων και η τρίτη συστάδα με το μικρότερο μέγεθος τίτλων, να έχει 413.08 σύνολο αναφορών, έχοντας 6.08 μέσο πλήθος λέξεων. Το δείγμα δεδομένων φαίνεται, εκτός από 2 έγγραφα (τα 2 πράσινα ακραία σημεία), να μην διαθέτει πάνω από 20 περίπου και πάνω πλήθους λέξεων. Οι 2 πρώτες συστάδες δείχνουν ότι τα έγγραφα με μήκος κάτω των 10 λέξεων, να αναφέρονται αρκετά συχνά, όμως η πρώτη συστάδα εκπροσωπεί και έγγραφα με χαμηλό αριθμό παραπομπών, ενώ ταυτόχρονα ο τίτλος αυτών αποτελείται από μικρό πλήθος λέξεων. Συνοψίζοντας, όπως έχουμε αναφέρει, ο τίτλος των εγγράφων αποτελεί έναν από τους πολλούς παράγοντες που επηρεάζουν την απήχηση και την αναφορά των εγγράφων, αλλά σαν παράγοντας από μόνος του, δεν είναι ικανός να διαδραματίσει σημαντικό ρόλο. Έτσι λοιπόν, οι συγγραφείς θα πρέπει να υπολογίσουν και να συνδυάσουν αρκετούς παράγοντες, με απώτερο

σκοπό τη συγγραφή και την έκδοση εργασιών που θα έχουν μεγάλο αντίκτυπο στην ευρεία επιστημονική κοινότητα. Βέβαια, το κάθε σύνολο δεδομένων που χρησιμοποιείται για τη διεξαγωγή αντίστοιχων πειραμάτων, διέπεται από τη δική του ποικιλομορφία.

5.3 Μοντέλο Encoder-Decoder LSTM

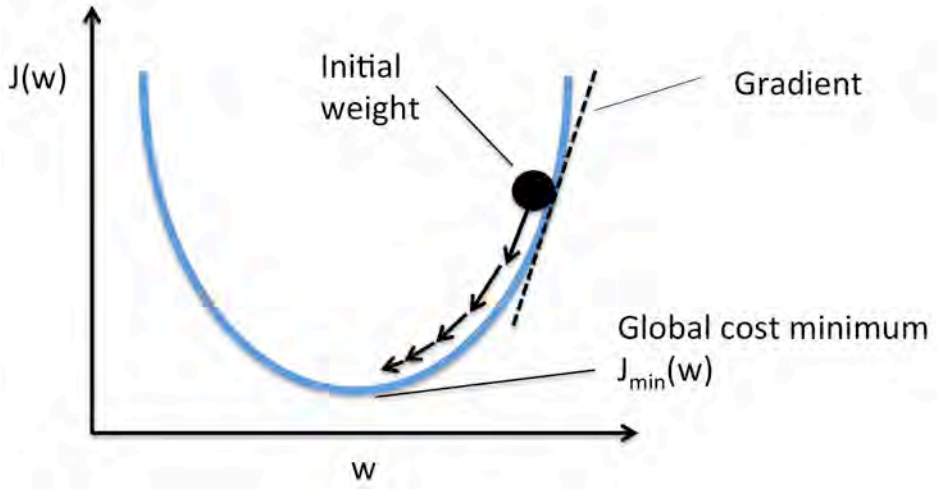
5.3.1 Τα Δεδομένα και οι Λεπτομέρειες Εκπαίδευσής του

Στην παρούσα εργασία, οι χρονοσειρές που θα συναντήσουμε, αφορούν το πλήθος των αναφορών που έλαβε το σύνολο των εγγράφων, κάθε χρονιά ξεχωριστά, από το 1900 μέχρι και το 2009. Συγκεκριμένα, για κάθε έγγραφο μετρήθηκε από τη χρονιά δημοσίευσής του μέχρι και το 2009, το σύνολο των παραπομπών που έλαβε από άλλα έγγραφα, τα οποία δημοσιεύτηκαν μετά από τη χρονιά δημοσίευσης του υπό εξέταση εγγράφου.

Για την εκπαίδευση του νευρωνικού δικτύου, οι χρονιές που ήταν οι πιο σημαντικές και αρκετές για να πετύχει σύγκλιση το νευρωνικό, ήταν από το 1995 μέχρι και το 2009. Επίσης, μετά από αφαίρεση κάποιων χρονολογικών σειρών που ήταν ίδιες ή κάποιων χρονοσειρών που αποτελούνταν από μηδενικά και άλλων όπου το άθροισμα των αναφορών σε όλες τις χρονιές δεν ξεπερνούσε τις 10, το σύνολο των χρονοσειρών που αξιοποιήθηκε για την εκπαίδευση του νευρωνικού ήταν 3000. Το 20% των δεδομένων, αποτέλεσε το σύνολο των δεδομένων δοκιμής και το υπόλοιπο αποτέλεσε το σύνολο εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από διάνυσμα εισόδου μεγέθους 12 και από διάνυσμα εξόδου μεγέθους 3, καθώς θέλουμε το νευρωνικό μας να προβλέπει τις αναφορές των εγγράφων τα επόμενα 3 χρόνια. Το διάνυσμα εξόδου αποτελεί τις ετικέτες του νευρωνικού, δηλαδή τις επιθυμητές για έξοδο τιμές. Τα δεδομένα εκπαίδευσης χωρίστηκαν με τη σειρά τους, σε δεδομένα εκπαίδευσης και δεδομένα επικύρωσης κατά το ίδιο ποσοστό, δηλαδή τα δεδομένα επικύρωσης είναι το 20% των αρχικών δεδομένων εκπαίδευσης. Το σύνολο των δεδομένων επικύρωσης, παρέχει μία αμερόληπτη αξιολόγηση ενός μοντέλου που ταιριάζει στο σύνολο δεδομένων εκπαίδευσης, ρυθμίζοντας παράλληλα τις υπερ-παραμέτρους του, ενώ το σύνολο των δεδομένων δοκιμής χρησιμοποιείται για να παρέχει μία αμερόληπτη αξιολόγηση του τελικού μοντέλου που ταιριάζει στο σύνολο δεδομένων εκπαίδευσης.

Συνήθως, τα νευρωνικά δίκτυα εκπαιδεύονται χρησιμοποιώντας τη στοχαστική κάθοδο κλίσης (Stochastic Gradient Descent) και απαιτούν να οριστεί κάποια συνάρτηση κόστους που ονομάζεται αλλιώς αντικειμενική συνάρτηση ή συνάρτηση απώλειας. Η κλίση αναφέρεται σε βαθμίδα σφάλματος που αφορά τις προβλέψεις. Η διαδικασία χαρτογράφησης των νευρωνικών δικτύων B.M, του συνόλου εισόδων σ' ένα σύνολο εξόδων, απαιτεί τα βέλτιστα δυνατά βάρη. Για να πραγματοποιηθεί αυτό, το πρόβλημα της μάθησης, εκτυλίσσεται σε πρόβλημα αναζήτησης ή βελτιστοποίησης, χρησιμοποιώντας κάποιον αλγόριθμο. Αυτός ο αλγόριθμος, χρησιμοποιείται για να εντοπίσει τα σύνολα βαρών που μπορούν να χρησιμοποιηθούν από το μοντέλο για να κάνει καλές προβλέψεις. Τα βάρη ανανεώνονται με τη χρήση της μεθόδου Error Back Propagation. Ο αλγόριθμος κλίσης στοχεύει στην τροποποίηση βαρών, με απώτερο σκοπό η επόμενη αξιολόγηση να επιφέρει μείωση στο σφάλμα, γεγονός που δείχνει ότι ο αλγόριθμος βελτιστοποίησης "κατευθύνει" προς τα κάτω την κλίση σφάλματος. Το σφάλμα μεταδίδεται στα προηγούμενα στρώματα, όπου χρησιμοποιεί-

ται για την τροποποίηση των βαρών (και της μεροληψίας), με απώτερο σκοπό την ελαχιστοποίησή του. Η επιλογή της συνάρτησης απώλειας συνδέεται άμεσα με την επιλογή της συνάρτησης ενεργοποίησης. Στο Σχ. 5.11, παρουσιάζεται η μέθοδος SGD, όπου η διαδικασία ξεκινάει από τα αρχικά βάρη και στοχεύει στην αναζήτηση ενός ολικού ελάχιστου σημείου κόστους.



Σχήμα 5.11: Απεικόνιση της στοχαστικής καθόδου κλίσης (Πηγή: [26])

Στο πρόβλημα μας, χρησιμοποιήθηκε ως μέτρο αξιολόγησης του μοντέλου, η συνάρτηση απώλειας που υπολογίζει το ριζικό του μέσου αθροίσματος των τετραγωνικών σφαλμάτων (διαφορών), μεταξύ των προβλεπόμενων τιμών του μοντέλου και των πραγματικών τιμών του προβλήματος. Η συνάρτηση αυτή είναι η Root Mean of Squared Error (RMSE). Όσο μικρότερες τιμές παράγει, τόσο το καλύτερο. Το αποτέλεσμα της είναι πάντα θετικό και ως ένδειξη βελτιστοποίησης του μοντέλου, θεωρείται η μείωση της τιμής της, με το ιδανικότερο να ισούται με το 0. Αποτελεί τη συνάρτηση στην οποία εφαρμόζεται ο αλγόριθμος βελτιστοποίησης Adam που θα αναλύσουμε παρακάτω. Επίσης, ως επιπλέον μέτρο αξιολόγησης του μοντέλου, χρησιμοποιήθηκε η συνάρτηση R^2, η οποία ονομάζεται συντελεστής προσδιορισμού. Η R^2 μετράει τη συσχέτιση μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Όσο μεγαλύτερη τιμή παράγει, τόσο μεγαλύτερη ακρίβεια επιδεικνύει το μοντέλο. Οι εξισώσεις 5.2 και 5.3 αποτελούν τον μαθηματικό τύπο της RMSE και της R^2 αντίστοιχα.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}, \tag{5.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5.3}$$

όπου n είναι ο αριθμός των δεδομένων, y_i είναι η i-οστή παρατηρούμενη τιμή των δεδομένων, x_i είναι η i-οστή τιμή της επεξηγηματικής μεταβλητής, $f(x_i)$ η προβλεπόμενη τιμή από την προσαρμογή της x_i στο μοντέλο και $\bar{y}$ ο μέσος όρος όλων των y_i.

Ο αλγόριθμος βελτιστοποίησης που εφαρμόστηκε, είναι ο αλγόριθμος Adam, όπως αναφέραμε προηγουμένως, με ρυθμό μάθησης 0.001. Αυτός ο αλγόριθμος αντικατέστησε την τυπική διαδικασία SGD που περιγράψαμε, για να ενημερώνει τα βάρη επαναληπτικά, με βάση τα δεδομένα εκπαίδευσης. Η SGD, διατηρεί ένα μοναδικό ρυθμό μάθησης για όλες τις ενημερώσεις βάρους και δεν αλλάζει κατά τη διάρκεια της εκπαίδευσης. Στον αλγόριθμο Adam, ο ρυθμός μάθησης διατηρείται για κάθε βάρος του δικτύου (παράμετρος) και προσαρμόζεται ξεχωριστά ως μάθηση που εκτυλίσσεται. Ο αλγόριθμος υπολογίζει ξεχωριστά τους προσαρμοστικούς ρυθμούς μάθησης, για διαφορετικές παραμέτρους από τις εκτιμήσεις των μέσων όρων της πρώτης και της δεύτερης στιγμής των κλίσεων. Η μέθοδος είναι απλή στην εφαρμογή, αποτελεσματική, έχει ελάχιστες απαιτήσεις μνήμης και είναι κατάλληλη για μεγάλα προβλήματα δεδομένων ή / και παραμέτρων. Επίσης, είναι κατάλληλη για προβλήματα με πολύ θορυβώδη ή / και αραιά κλίματα.

Ο αριθμός περιόδων χρόνου έχει οριστεί σε 80. Συγκεκριμένα, ο αριθμός αυτός που αφορά συγκεκριμένη περίοδο χρόνου, είναι όταν ολόκληρο το σύνολο δεδομένων εκπαίδευσης μεταβιβάζεται στο νευρωνικό δίκτυο. Επειδή όμως, αυτός ο αριθμός είναι μεγάλος για να τροφοδοτηθεί αμέσως με δεδομένα το μοντέλο, χωρίζεται σε μικρότερες παρτίδες. Έτσι, το μέγεθος παρτίδας έχει οριστεί σε 200.

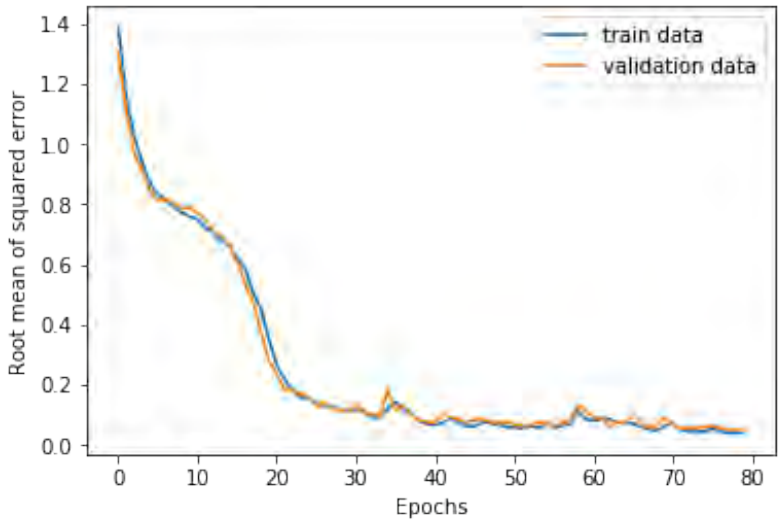
Έτσι, μετά το πέρας της διαδικασίας εκπαίδευσης του νευρωνικού μας δικτύου, έχει προκύψει ένα μοντέλο το οποίο είναι ικανό να προβλέπει τις παραπομπές της επόμενης τριετίας ενός εγγράφου, με βάση τα ιστορικά δεδομένα των πρώτων 12 ή και παραπάνω χρόνων αναφοράς. Επιπλέον, προβλέποντας τον αριθμό των παραπομπών ενός εγγράφου, είναι εφικτή η αξιολόγηση του μελλοντικού αντίκτυπου των συγγραφέων, με πιθανές εφαρμογές στην πρόσληψη αυτών, καθώς και τη χορήγηση βραβείων και κονδυλίων στις αντίχοιχες επιστημονικές έρευνες. Βέβαια, όπως έχουμε αναφέρει ξανά, η πρόβλεψη μελλοντικών αναφορών συνοδεύεται από πολλούς παράγοντες και δεν επιτυγχάνεται πλήρως από την ανάπτυξη ενός τέτοιου μοντέλου. Χαρακτηριστικό παράδειγμα είναι ότι κάποια έγγραφα για πολλά χρόνια παραμένουν απαρατήρητα, ενώ κάποια στιγμή προσελκύουν μεγάλη προσοχή, ενώ υπάρχουν άλλα έγγραφα που ήταν ευρέως αναγνωρίσιμα και αναφερόμενα και η απήχησή τους μειώθηκε δραματικά, λόγω της ανάπτυξης νέων πιθανολογουμένως, καλύτερων μεθόδων.

5.3.2 Παρουσίαση του Ιστορικού Εκπαίδευσής του

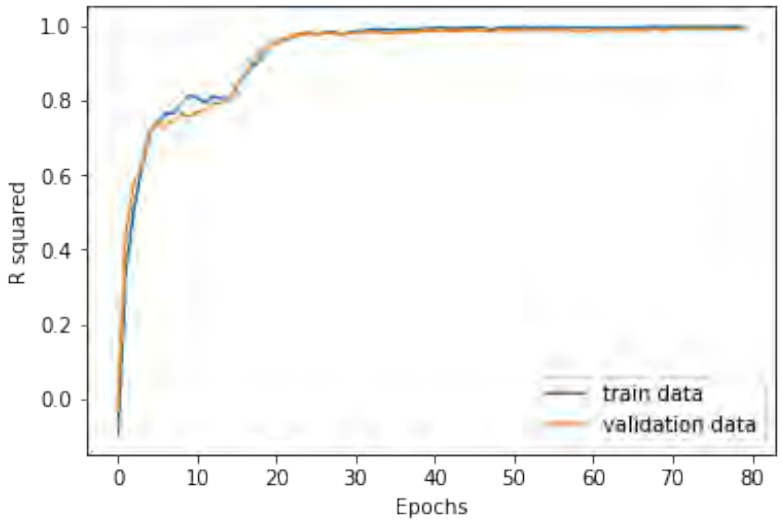
Μπορούμε να μαθαίνουμε πολλά πράγματα για τα νευρωνικά δίκτυα και για τα μοντέλα της Β.Μ, παρατηρώντας την απόδοσή τους με την πάροδο του χρόνου, καθ' όλη τη διάρκεια της εκπαίδευσης. Υπάρχει η δυνατότητα καταγραφής μετρήσεων εκπαίδευσης για κάθε συγκεκριμένη περίοδο χρόνου, καθώς και η δυνατότητα επίγνωσης της απώλειας και της ακρίβειας για το σύνολο των δεδομένων επικύρωσης, εφόσον αυτό έχει οριστεί. Η διαδικασία βελτιστοποίησης αφορά τις συναρτήσεις απώλειας που έχουν καθοριστεί, ενώ αυτές ή και συναρτήσεις όπως η R^2 που αναφέραμε, αποτελούν τη βάση για την εξέταση της απόδοσης του μοντέλου.

Έχοντας στη διάθεση μας το ιστορικό εκπαίδευσης του μοντέλου, η δημιουργία γραφημάτων καθίσταται εφικτή. Είναι μία διαδικασία, μέσω της οποίας μπορούμε να βλέπουμε την ταχύτητα σύγκλισης του μοντέλου σε συγκεκριμένες περιόδους χρόνου (κλίση), να δούμε αν το μοντέλο

μπορεί να έχει ήδη συγκλίνει ή να έχει πραγματοποιηθεί υπερβολική μάθηση των δεδομένων εκπαίδευσης. Δημιουργήθηκαν 2 γραφήματα που αφορούν ολόκληρη τη διαδικασία της εκπαίδευσης. Το ένα γράφημα αφορά τη συνάρτηση απώλειας RMSE και το άλλο γράφημα αφορά τη συνάρτηση R^2 που χρησιμοποιήθηκε ως ένα επιπλέον μέτρο για την αξιολόγηση του μοντέλου. Τα 2 γραφήματα αξιολόγησης σχετίζονται με τα σύνολα δεδομένων εκπαίδευσης και επικύρωσης και απεικονίζονται παρακάτω. Στο Σχ. 5.12, έχουμε τη συνάρτηση αξιολόγησης του μοντέλου (συνάρτηση απώλειας), την RMSE, για 80 epochs, παρατηρώντας να μειώνεται σε αρκετά μεγάλο βαθμό, πράγμα που σημαίνει ότι το μοντέλο εκπαιδεύεται σωστά και στο Σχ. 5.13, έχουμε τη συνάρτηση R^2, όπου παρατηρούμε ότι οι προβλεπόμενες τιμές τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα επικύρωσης, έχουν ισχυρή συσχέτιση, με τη συνάρτηση να φτάνει κοντά στο 1.



Σχήμα 5.12: Γράφημα RMSE για δεδομένα εκπαίδευσης και επικύρωσης



Σχήμα 5.13: Γράφημα R^2 για δεδομένα εκπαίδευσης και επικύρωσης

Το μοντέλο μας δεν χαρακτηρίζεται από υπο-προσαρμογή, καθώς ένα μοντέλο που διέπεται από υπο-προσαρμογή έχει καλές επιδόσεις στο σύνολο δεδομένων εκπαίδευσης και κακές στα δεδομένα επικύρωσης, κάτι που δεν παρατηρείται με βάση την καμπύλη των 2 γραφημάτων. Επιπρόσθετα, το μοντέλο μας δεν μπορεί να χαρακτηριστεί από υπερ-προσαρμογή, καθώς το φαινόμενο της υπερ-προσαρμογής, συμβαίνει όταν η καμπύλη των δεδομένων εκπαίδευσης συνεχίζει να βελτιώνεται, ενώ η καμπύλη των δεδομένων επικύρωσης βελτιώνεται ως ένα σημείο και μετά αρχίζει να υποβαθμίζεται. Επίσης, η καμπύλη του γραφήματος της R^2, αγγίζει τιμές κοντά στο 1. Έτσι, το μοντέλο που αναπτύχθηκε, εκπαιδεύτηκε σωστά και προβλέπει αρκετά επιτυχώς, δεδομένα που δεν έχει "συναντήσει" προηγουμένως. Οι κλίσεις των δεδομένων εκπαίδευσης και επικύρωσης στα γραφήματα, σταθεροποιούνται περίπου στο ίδιο σημείο, γεγονός που δείχνει ότι το μοντέλο είναι αρκετά καλό. Τέλος, το μοντέλο μας παρουσιάζει ακρίβειες πρόβλεψης στα δεδομένα εκπαίδευσης, της τάξης του 98%, 99% και 99% στις 3 χρονιές πρόβλεψης αντίστοιχα.

5.3.3 Παράδειγμα Πρόβλεψης Μελλοντικών Αναφορών

Η χρονοσειρά εισόδου, αναφέρεται στο πλήθος αναφορών που έλαβε ένα έγγραφο και πρέπει να αποτελείται από τουλάχιστον 12 χρονιές, ενώ η παραγόμενη χρονοσειρά εξόδου, αναφέρεται στο πλήθος των αναφορών που θα λάβει το έγγραφο τις επόμενες 3 χρονιές, από την τελευταία χρονιά αναφοράς και έπειτα. Στον Πίνακα 5.3, παρουσιάζονται οι προβλεπόμενες αναφορές για 5 έγγραφα. Ως ακολουθία εισόδου είναι 5 χρονοσειρές μεγέθους 12 και ως ακολουθία εξόδου είναι 5 προβλεπόμενες χρονοσειρές μεγέθους 3. Στην προκειμένη περίπτωση, οι χρονοσειρές είχαν όλες μήκος 12. Το νευρωνικό που αναπτύξαμε όμως, μπορεί να προβλέπει και για χρονοσειρές, με μήκος μεγαλύτερο του 12. Οι τιμές με έντονο μαύρο χρώμα, είναι οι προβλεπόμενες αναφορές για την κάθε χρονοσειρά αντίστοιχα.

18	46	69	93	105	133	144	167	194	153	140	129	**143**	**94**	**108**
2	6	3	4	10	11	4	4	8	3	3	2	**7**	**3**	**3**
34	55	57	63	59	49	57	67	66	71	68	66	**64**	**46**	**50**
8	9	13	11	12	17	7	5	7	7	4	4	**7**	**8**	**4**
6	8	10	13	13	17	17	23	21	19	17	15	**20**	**20**	**18**

Πίνακας 5.3: Προβλέψεις αναφορών για 5 χρονοσειρές

5.3.4 Τροποποίηση των Παραμέτρων του

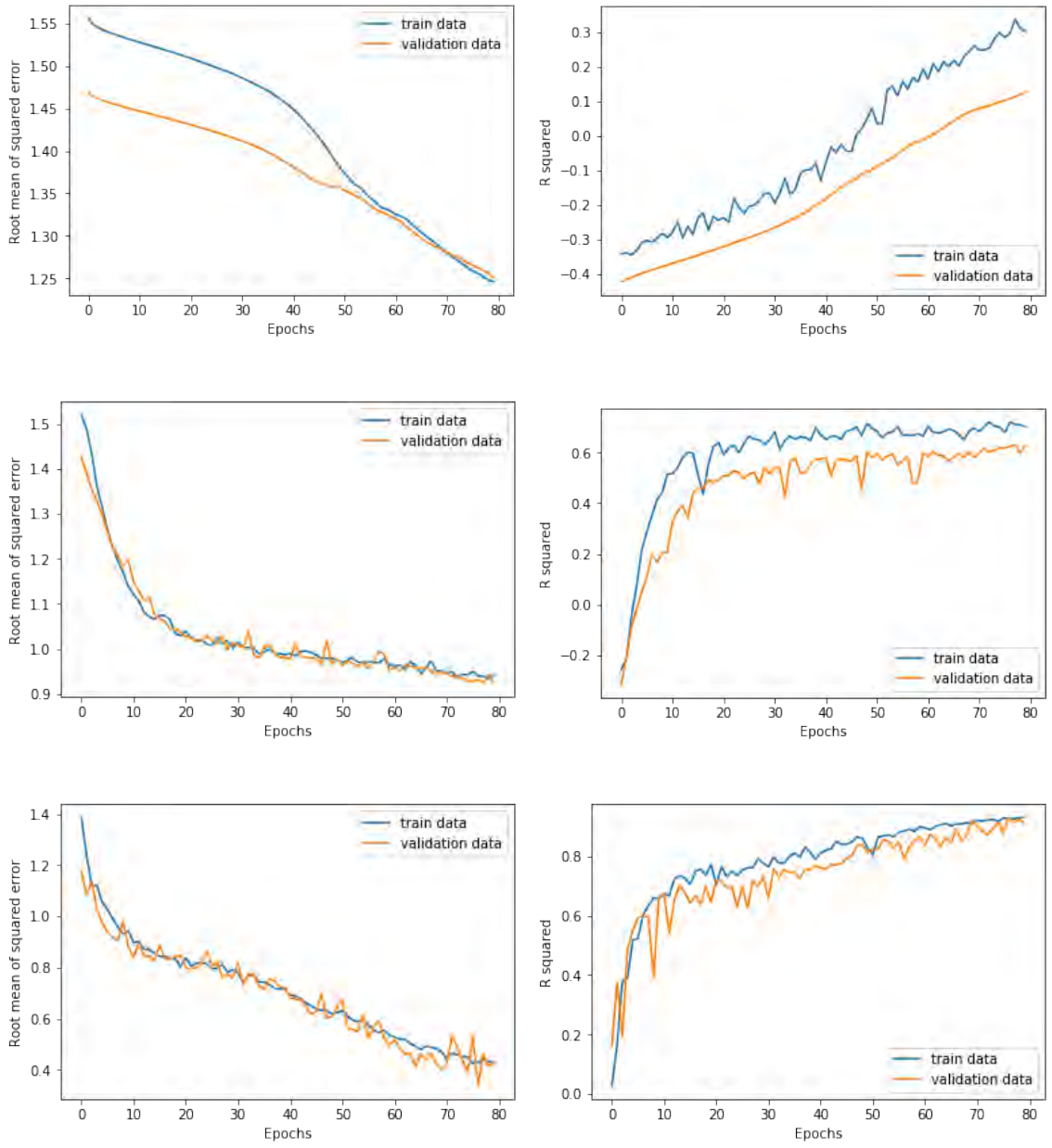
Οι παράμετροι που επιλέξαμε να τροποποιήσουμε ήταν οι αλγόριθμοι βελτιστοποίησης που αφορούν τη συνάρτηση απώλειας RMSE του μοντέλου. Χρησιμοποιήθηκαν οι μέθοδοι SGD και RMSProp (Root Mean Square Propagation), με διαφορετικούς ρυθμούς μάθησης και παρακάτω παρατίθενται τα γραφήματα RMSE και R^2 που προέκυψαν.

Αναλυτικότερα, στο Σχ. 5.14, βλέπουμε τα γραφήματα των 2 συναρτήσεων και παρατηρούμε

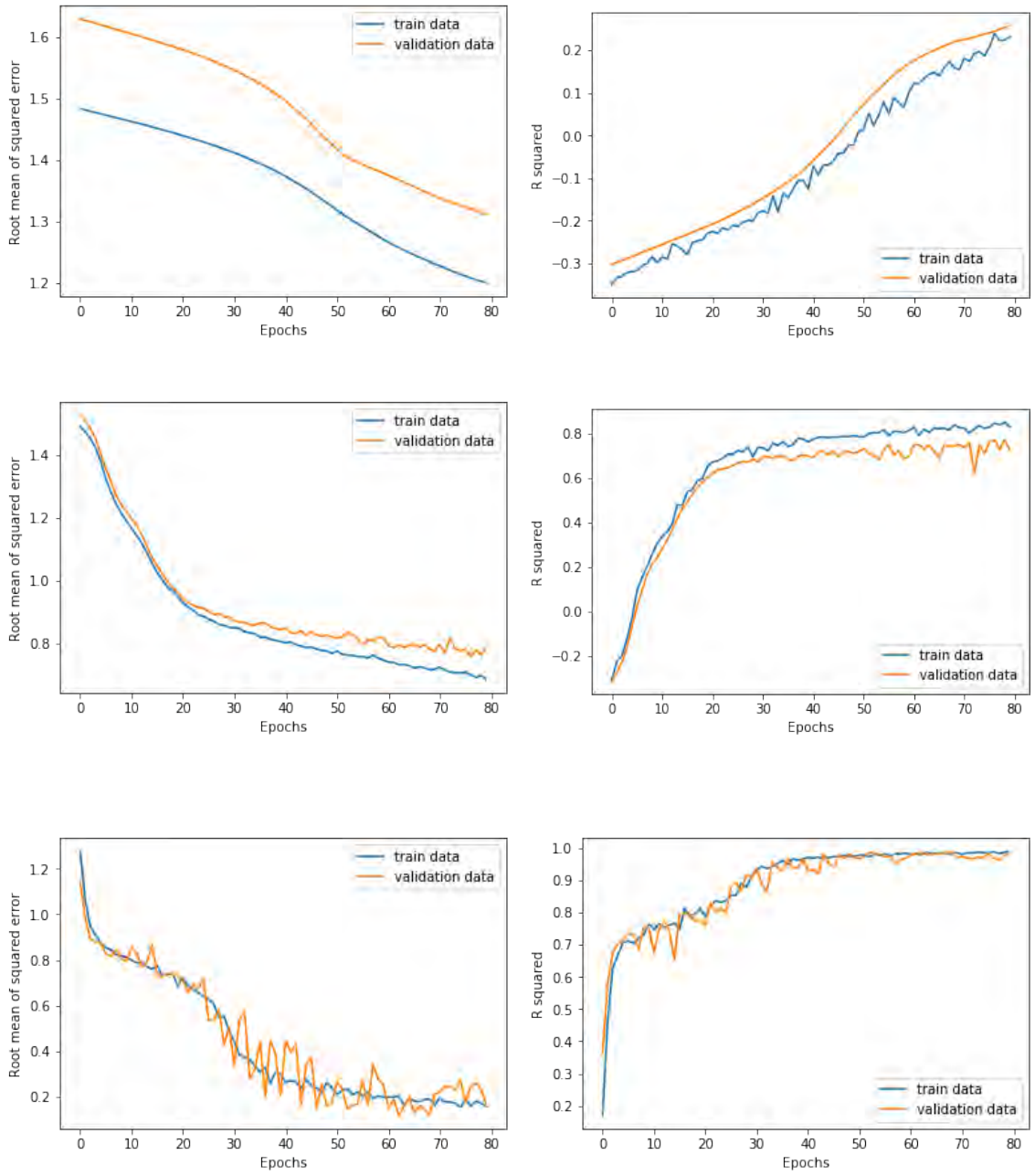
πως η μέθοδος SGD για 80 επαναλήψεις και για χαμηλούς ρυθμούς μάθησης, δεν επιδεικνύει καλά αποτελέσματα (σύγκλιση) όπως ο αλγόριθμος Adam που χρησιμοποιήσαμε στο νευρωνικό δίκτυο. Για μεγαλύτερους ρυθμούς μάθησης, παρουσιάζει καλύτερα αποτελέσματα. Αυξάνοντας ακόμα περισσότερο τον ρυθμό μάθησης για τη μέθοδο SGD, η γραφική παράσταση της απόδοσής της καθίσταται ανέφικτη, καθώς η R^2 και η RMSE, ξεκινάνε από μεγάλες αρνητικές και θετικές τιμές αντίστοιχα.

Στη συνέχεια, εφαρμόστηκε ο αλγόριθμος RMSprop για τον ίδιο αριθμό επαναλήψεων (80) και με διαφορετικούς ρυθμούς μάθησης. Ο αλγόριθμος RMSprop, είναι ένας αλγόριθμος μάθησης βασισμένος στην κάθοδο κλίσης και συνδυάζει 2 μεθόδους βελτιστοποίησης, την Adagrad και την Adadelta. Παρατηρώντας τα γραφήματα στο Σχ. 5.15, παρατηρούμε ότι για χαμηλούς ρυθμούς μάθησης, δεν επιδεικνύει καλά αποτελέσματα. Για μεγαλύτερες τιμές του ρυθμού μάθησης, παρουσιάζει καλύτερα αποτελέσματα, αλλά δεν συγκλίνει το ίδιο γρήγορα με τη μέθοδο που χρησιμοποιήσαμε, ενώ τα δεδομένα επικύρωσης έχουν αρκετές διακυμάνσεις στο γράφημα της RMSE. Μεγαλώνοντας περισσότερο την τιμή του ρυθμού μάθησης, το μοντέλο δεν παρουσιάζει καλά αποτελέσματα, έχοντας μεγάλες απώλειες και κακή συσχέτιση μεταξύ των πραγματικών και των προβλεπόμενων τιμών.

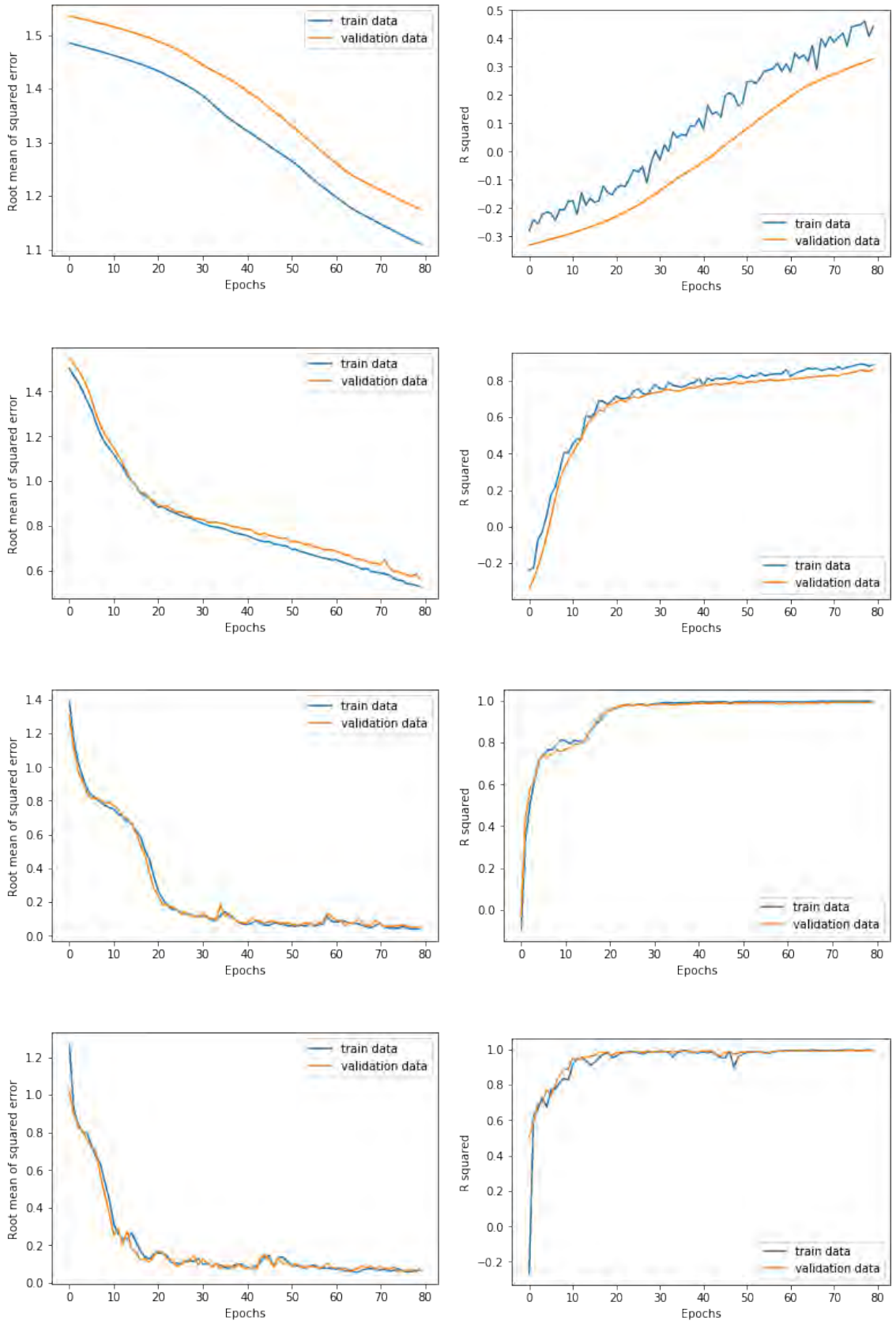
Τέλος, στο Σχ. 5.16 απεικονίζεται η γραφική παράσταση των 2 συναρτήσεων για διαφορετικούς ρυθμούς μάθησης του αλγορίθμου Adam που εφαρμόσαμε στο τελικό νευρωνικό δίκτυο. Για τιμές μεγαλύτερες ή ίσες του 0.1, το μοντέλο χαρακτηρίζεται από απώλειες μεγάλης κλίμακας και αυτός ήταν και ο λόγος που δεν απεικονίζονται τα γραφήματα των συναρτήσεων για αυτές τις τιμές. Για τιμές 0.01, 0.001, παρουσιάζει περίπου την ίδια σύγκλιση. Έτσι, κρατήσαμε ως αλγόριθμο βελτιστοποίησης τον Adam, με ρυθμό μάθησης 0.001, καθώς στα δεδομένα εκπαίδευσης φαίνεται η συσχέτιση μεταξύ πραγματικών και προβλεπόμενων τιμών να παραμένει πιο σταθερή σε υψηλές τιμές. Για τιμές μικρότερες ή ίσες του 0.0001, το μοντέλο αδυνατεί να συγκλίνει στον ίδιο αριθμό επαναλήψεων.



Σχήμα 5.14: Γραφήματα RMSE - R^2 με SGD για ρυθμούς μάθησης: 0.001, 0.01, 0.1



Σχήμα 5.15: Γραφήματα RMSE - R^2 με RMSprop για ρυθμούς μάθησης: 0.00001, 0.0001, 0.001



Σχήμα 5.16: Γραφήματα RMSE - R^2 με Adam για ρυθμούς μάθησης: 0.00001, 0.0001, 0.001, 0.01

5.4 Λογισμικά

5.4.1 Python & Βιβλιοθήκες

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για τη μελέτη και την ανάπτυξη των μοντέλων είναι η Python. Η Python είναι γλώσσα υψηλού επιπέδου, με αρκετές εφαρμογές και χρησιμοποιείται αρκετά συχνά στον τομέα της ανάλυσης δεδομένων. Η ανάλυση αυτών των δεδομένων, συντέλεσε στη δημιουργία πολλών πλατφόρμων που χρησιμοποιούν αυτή τη γλώσσα. Χρησιμοποιείται ευρέως από μεγάλους οργανισμούς, λόγω των πολλών παραδειγμάτων προγραμματισμού της.

Η Python χαρακτηρίζεται από μία μεγάλη γκάμα βιβλιοθηκών, από αυτόματη διαχείριση μνήμης και από δυναμικά χαρακτηριστικά. Μερικές από τις βασικές βιβλιοθήκες που χρησιμοποιήθηκαν για την εκπόνηση των πειραμάτων, είναι η Pandas, η οποία αποτελείται από ένα σύνολο συναρτήσεων για μελέτη και παρουσίαση συνόλων δεδομένων, η NumPy που περιέχει συναρτήσεις για μαθηματικές πράξεις αλλά και για μελέτη και διαχείριση πινάκων αυτού του τύπου, καθώς και η Matplotlib, η οποία χρησιμοποιείται για τη δημιουργία και αναπαράσταση γραφικών παραστάσεων. Κάποιες επιπλέον βιβλιοθήκες που συντέλεσαν στη διεξαγωγή πειραμάτων, είναι οι ακόλουθες:

- Η **pyLDAvis** είναι μία από τις βιβλιοθήκες της Python και χρησιμοποιείται για τη διαδραστική απεικόνιση της μοντελοποίησης θεμάτων. Η βιβλιοθήκη αυτή, έχει σχεδιαστεί για να βοηθάει τους χρήστες να ερμηνεύουν τα θέματα που προέκυψαν από το LDA μοντέλο, το οποίο είναι προσαρμοσμένο στο σύνολο των δεδομένων κειμένου.

- Η **Scikit-learn** είναι βιβλιοθήκη γενικού σκοπού και αφορά τη Μ.Μ. Υποστηρίζει πολλούς αλγορίθμους και προσφέρει πολλά χαρακτηριστικά σχετικά με την ταξινόμηση κειμένου, την παλινδρόμηση και τα μοντέλα συσταδοποίησης. Είναι από τις πιο ευκολόχρηστες βιβλιοθήκες και παρέχει πολλούς οδηγούς για την καλύτερη κατανόησή της.

- Η **NLTK** βιβλιοθήκη εστιάζει στον τομέα της Ε.Φ.Γ και είναι πολύ χρήσιμη στην ταξινόμηση κειμένου. Αποτελείται από ένα σύνολο χρήσιμων εργαλείων, τα οποία καθιστούν μία μηχανή ικανή να "κατανοεί" κείμενα. Κάποια παραδείγματα κατανόησης κειμένου είναι η αναγνώριση των μερών του λόγου και ο διαχωρισμός των λέξεων.

- Η **spaCy** αποτελεί τον καλύτερο τρόπο προεπεξεργασίας κειμένου που προορίζεται για σκοπούς της Β.Μ. Δίνει τη δυνατότητα για την κατασκευή πολύπλοκων γλωσσικών μοντέλων στατιστικής για ποικιλία προβλημάτων της Ε.Φ.Γ. Συντέλεσε στη μοντελοποίηση θεμάτων που πραγματοποιήσαμε, δίνοντας τη δυνατότητα για επιλογή λέξεων, με βάση τα επιθυμητά μέρη του λόγου.

- Η **Gensim** βιβλιοθήκη, μία δημοφιλής βιβλιοθήκη της Μ.Μ, αφορά την ομαδοποίηση κειμένου. Στόχος της είναι η θεματική ανάλυση μεγάλης κλίμακας αδόμητων δεδομένων. Αποτέλεσε εργαλείο για την αποτελεσματικότερη εξαγωγή θεμάτων που πραγματοποιήθηκε. Γενικότερα, όλες οι προαναφερόμενες βιβλιοθήκες, αποτέλεσαν τη βάση για το πείραμα

της μοντελοποίησης θεμάτων, εκτός της Scikit-learn που αξιοποιήθηκε και στα πειράματα συσταδοποίησης και στο μοντέλο πρόβλεψης αναφορών.

5.4.2 Keras & Tensorflow

Όσον αφορά τη δημιουργία του νευρωνικού δικτύου που προβλέπει μελλοντικές αναφορές, χρησιμοποιήθηκε ένα Application Programming Interface (API), το Keras. Είναι απλή, υψηλού επιπέδου βιβλιοθήκη νευρωνικών δικτύων που χρησιμοποιείται για την κατασκευή και την εκπαίδευση μοντέλων της Β.Μ. Το παραπάνω API είναι φιλικό προς τον χρήστη, εύκολα επεκτάσιμο, τα μοντέλα του δεν απαιτούν πολλούς περιορισμούς και κατασκευάζονται με σύνδεση διαμορφωμένων δομικών στοιχείων. Λειτουργεί ως "περιτύλιγμα" στη βιβλιοθήκη TensorFlow ή στη Theano και όλα μαζί αποτελούν τα βασικά πλαίσια της Β.Μ. Για τη λειτουργία του Keras, απαιτείται η ύπαρξη μίας από τις 2 προαναφερόμενες βιβλιοθήκες. Στην εργασία χρησιμοποιήθηκε η TensorFlow που είναι μία βιβλιοθήκη ανοικτού κώδικα και χρησιμοποιείται για την κατασκευή και την παραγωγή μοντέλων Β.Μ.

5.4.3 Anaconda

Αναφέραμε προηγουμένως, ότι υπάρχουν πολλές πλατφόρμες που χρησιμοποιούν την Python. Μία τέτοια είναι και η πλατφόρμα Anaconda που είναι μία δωρεάν πλατφόρμα, αποτελούμενη από αρκετά προγράμματα για τη μελέτη δεδομένων. Αυτή η πλατφόρμα διαθέτει 2 γλώσσες προγραμματισμού, οι οποίες είναι η Python και η R. Είναι ευρέως γνωστή και χρησιμοποιείται αρκετά συχνά, με δυνατό της σημείο την ευελιξία στην επίλυση προβλημάτων σε πολλούς τομείς. Η εφαρμογή που χρησιμοποιήσαμε για την κατασκευή των μοντέλων και αποτελεί ένα από τα προγράμματα της Anaconda, είναι το Jupyter Notebook.

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη και Συμπεράσματα

Συνοψίζοντας όλα τα παραπάνω και βλέποντας τα αποτελέσματα των διεξαχθέντων πειραμάτων, παρατηρούμε ότι υπάρχουν τεχνικές της Μ.Μ που συντελούν στην ανάλυση της επιστήμης και την εξαγωγή πληροφοριών στον τομέα της επιστημονομετρίας. Αρχικά, ο αλγόριθμος LDA που χρησιμοποιήθηκε για τη μοντελοποίηση θεμάτων, είναι αρκετά καλός και αποτελεσματικός σε σύγκριση με άλλες μεθόδους, όπως οι μέθοδοι LSI και pLSI που αναφέραμε στη Βιβλιογραφική Ανασκόπηση 2.1, όπου περιγράψαμε αποτελέσματα άλλων ερευνών. Έτσι λοιπόν, τροποποιώντας παραμέτρους του αλγορίθμου, και συγκρίνοντας τα αποτελέσματα από τα μετρικά που αξιοποιήθηκαν για την αξιολόγησή του, παρατηρήσαμε ότι η δημιουργία θεμάτων που πραγματεύονται τα έγγραφα με τον αλγόριθμο LDA, λειτουργεί αρκετά καλά.

Όσον αφορά το κομμάτι της συσταδοποίησης εγγράφων με συντεταγμένες το μέγεθος των τίτλων τους και τις συνολικές παραπομπές που έλαβαν τη χρονική περίοδο 1900-2009, χρησιμοποιήθηκε ο αλγόριθμος K-Means. Αυτός ο αλγόριθμος είναι απλός στην εφαρμογή και αποτελεί μία από τις πιο κοινές διερευνητικές τεχνικές ανάλυσης δεδομένων που αξιοποιούνται στην απόκτηση "διαίσθησης", σχετικά με τη δομή των εξεταζόμενων δεδομένων. Σε αντίθεση με την εποπτευόμενη μάθηση όπου υπάρχουν μέτρα για την αξιολόγηση των μοντέλων, η διαδικασία συσταδοποίησης δεν έχει κάποια σταθερή μετρική αξιολόγησης. Ο K-Means δεν μαθαίνει από τα δεδομένα τον κατάλληλο αριθμό συστάδων που πρέπει να δημιουργήσει και ως εκ τούτου, δεν υπάρχει τρόπος για την εύρεση της βέλτιστης λύσης σχετικά με τον κατάλληλο αριθμό συστάδων. Μία από τις τεχνικές που εφαρμόζεται για την απόκτηση κάποιας "διαίσθησης" σχετικά με τον κατάλληλο αριθμό των συστάδων, χρησιμοποιήθηκε και στο δικό μας πείραμα και ονομάζεται τεχνική Elbow. Η τεχνική αυτή μας έδειξε, βασιζόμενη στο άθροισμα των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και πραγματικών τιμών, ότι 3 είναι οι καταλληλότερες συστάδες για την ομαδοποίηση των εγγράφων. Στη συσταδοποίηση εγγράφων που "μοιράζονται" παρόμοια θέματα, ο αριθμός των συστάδων δεν χρειάστηκε να ερευνηθεί, καθώς ισούται με τον αριθμό θεμάτων που "ανακάλυψε" η διαδικασία της μοντελοποίησης θεμάτων.

Τέλος, αναφορικά με την πρόβλεψη μελλοντικών παραπομπών που θα λάβει ένα έγγραφο μέσα στην επόμενη τριετία, μετά την τελευταία χρονιά που καταγράφεται το πλήθος των ληφθέντων

παραπομπών του, το τεχνητό νευρωνικό δίκτυο που αναπτύχθηκε ήταν τύπου LSTM Encoder-Decoder. Τα μέτρα αξιολόγησης (RMSE, R^2) που χρησιμοποιήθηκαν, σε συνδυασμό με την τροποποίηση παραμέτρων (αλγόριθμοι βελτιστοποίησης) του νευρωνικού δικτύου, τα αποτελέσματα των γραφημάτων αξιολόγησης έδειξαν ότι το μοντέλο μαθαίνει αρκετά καλά και είναι ικανό να προβλέπει, με αρκετά καλή ακρίβεια, τις μελλοντικές αναφορές των εγγράφων. Στα δεδομένα εκπαίδευσης, επέδειξε ακρίβεια της τάξης του 98%, 99%, και 99%, στις επόμενες 3 χρονιές πρόβλεψης αντίστοιχα.

6.2 Μελλοντικές Επεκτάσεις

Αναφορικά με τις μελλοντικές επεκτάσεις που θα μπορούσαν να γίνουν στην παρούσα διπλωματική, αυτές ξεκινάνε τόσο από τα επιστημονομετρικά δεδομένα όσο και από τα μοντέλα που αναπτύχθηκαν. Έτσι, μπορούμε να χρησιμοποιήσουμε δεδομένα που θα διαθέτουν περισσότερα δημοσιευμένα έγγραφα, τα οποία θα διέπονται και από περισσότερα γνωρίσματα. Η ύπαρξη περισσότερων γνωρισμάτων στα δεδομένα, θα συντελούσε στη διεξαγωγή περισσότερων πειραμάτων αλλά και στον εκτενέστερο εντοπισμό πληροφοριών. Τα επιστημονομετρικά δεδομένα της εκπονηθείσας διπλωματικής διέθεταν 7 γνωρίσματα. Όσον αφορά τώρα τα μοντέλα, η συγκέντρωση περισσότερων εγγράφων από μόνη της θα βοηθούσε στη καλύτερη εκπαίδευση αυτών και στη δημιουργία ποικίλων αποτελεσμάτων. Επίσης, ως μελλοντική επέκταση θα μπορούσε να αποτελέσει η εφαρμογή διαφορετικών μοντέλων ή / και διαφορετικών παραμέτρων αυτών που αφορούν τη μοντελοποίηση θεμάτων, τη συσταδοποίηση και την πρόβλεψη μελλοντικών αναφορών. Τέλος, έχει αποδειχθεί από τους επιστήμονες, ότι με περισσότερα στρώματα στο νευρωνικό δίκτυο, μπορούμε να πετύχουμε μεγαλύτερη ακρίβεια σε διάφορα προβλήματα. Έτσι, η παρούσα διπλωματική μπορεί να θεωρηθεί ως βάση για μελλοντικές εργασίες ή πειράματα σχετικά με την πρόβλεψη αναφορών για ένα σύνολο δημοσιευμένων εγγράφων, με την κατασκευή “βαθύτερων” δικτύων, για την παραγωγή ακόμη πιο περίπλοκων και ισχυρών προβλέψεων.

Παράρτημα Ι

Κατηγοριοποίηση των Πηγών

Στο παράρτημα αυτό, κατηγοριοποιούμε τα είδη των πηγών που συμβουλευτήκαμε σε διάφορες κατηγορίες, με απώτερο σκοπό την καλύτερη οργάνωση και παρουσίαση αυτών. Οι κατηγορίες είναι οι εξής:

- Βιβλία Ξενόγλωσσα [29]
- Άρθρα σε επιστημονικά περιοδικά [4, 9, 10, 20, 25, 31, 34, 35, 38, 40, 44, 45, 48, 52, 60, 61]
- Άρθρα σε διεθνή επιστημονικά συνέδρια [5, 11, 12, 24, 33, 46, 47, 67, 68, 69]
- Ιστοσελίδες [1, 2, 3, 6, 7, 8, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 26, 27, 28, 30, 32, 36, 37, 39, 41, 42, 43, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 62, 64, 65, 66, 70, 71, 72]

Βιβλιογραφία

[1] A Gentle Introduction to k-fold Cross-Validation. https://machinelearningmastery.com/k-fold-cross-validation/. Ημερομηνία πρόσβασης: 23/5/2018.

[2] A Simple Introduction to Natural Language Processing. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32. Ημερομηνία πρόσβασης: 15/10/2018.

[3] About Train, Validation and Test Sets in Machine Learning. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7. Ημερομηνία πρόσβασης: 6/12/2017.

[4] Ali Abrishami και Sadegh Aliakbary. NNCP: A citation count prediction methodology based on deep neural network learning techniques. *CoRR*, abs/1809.04365, 2018.

[5] Sheena Angra και Sachin Ahuja. International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). Στο *Machine learning and its applications: A review*, Chirala, India, 2017.

[6] Article Level Metrics. https://sparcopen.org/our-work/article-level-metrics/.

[7] Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference? https://www.sumologic.com/blog/machine-learning-deep-learning/. Ημερομηνία πρόσβασης: 11/10/2018.

[8] Artificial Neural Networks with Net# in Azure ML Studio. https://naadispeaks.wordpress.com/2017/11/08/artificial-neural-networks-with-net-in-azure-ml-studio/. Ημερομηνία πρόσβασης: 8/11/2017.

[9] Kaveh Bastani, Hamed Namavari και Jeffrey Shaffer. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Syst. Appl.*, 127:256–271, 2019.

[10] David M. Blei, Andrew Y. Ng και Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] Jason Chuang, Christopher D. Manning και Jeffrey Heer. Termite: visualization techniques for assessing textual topic models. Στο *International Working Conference on Advanced Visual Interfaces, AVI 2012, Capri Island, Naples, Italy, May 22-25, 2012, Proceedings*, σελίδες 74–77, 2012.

[12] Jason Chuang, Christopher D. Manning και Jeffrey Heer. Termite: visualization techniques for assessing textual topic models. Στο *International Working Conference on Advanced Visual Interfaces, AVI 2012, Capri Island, Naples, Italy, May 22-25, 2012, Proceedings*, σελίδες 74–77, 2012.

[13] Citation Network Dataset. `https://aminer.org/citation`.

[14] Coding Deep Learning For Beginners. `https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d`. Ημερομηνία πρόσβασης: 25/7/2018.

[15] Cross-Validation Explained. `http://genome.tugraz.at/proclassify/help/pages/XV.html`.

[16] Data Mining vs. Machine Learning: What's The Difference? `https://www.import.io/post/data-mining-machine-learning-difference/`. Ημερομηνία πρόσβασης: 11/11/2019.

[17] Data Mining. What it is and why it matters. `https://www.sas.com/el_gr/insights/analytics/data-mining.html`.

[18] Datasets and Machine Learning. `https://skymind.ai/wiki/datasets-ml`.

[19] Deep Learning Explained in 7 Steps. `https://www.datadriveninvestor.com/2019/01/23/deep-learning-explained-in-7-steps/`. Ημερομηνία πρόσβασης: 23/1/2019.

[20] Boer Deng. Papers with shorter titles get more citations. *Nature*, 2015.

[21] Derivative of the Sigmoid function - a worked example. `http://ronny.rest/blog/post_2017_08_10_sigmoid/`. Ημερομηνία πρόσβασης: 10/8/2017.

[22] Eigenfactor vs. Impact Factor: How are They Different? `https://rushu.libguides.com/metrics`. Ημερομηνία πρόσβασης: 22/5/2018.

[23] Encoder-Decoder Long Short-Term Memory Networks. `https://machinelearningmastery.com/encoder-decoder-long-short-term-memory-networks/`. Ημερομηνία πρόσβασης: 23/8/2017.

[24] Dipesh Gautam, Nabin Maharjan, Rajendra Banjade, Jimba Lasang, Vasile Tamang και Rus . Long Short Term Memory based Models for Negation Handling in Tutorial Dialogues. Στο *The Thirty-First International Florida - Artificial Intelligence Research Society Conference (FLAIRS-31)*, Melbourne, Florida, USA, 2018.

[25] Sukhdev Singh Ghuman. Clustering Techniques- A Review. *International Journal of Computer Science and Mobile Computing*, abs/1809.04365:524–530, 2016.

[26] Gradient Descent and Stochastic Gradient Descent. http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/.

[27] Hierarchical Clustering. https://www.saedsayad.com/clustering_hierarchical.htm.

[28] How to Build A Data Set For Your Machine Learning Project. https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac.

[29] Judith Hurwitz και Daniel Kirsch. *Machine Learning For Dummies*. John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ, 2018.

[30] Intuitive Guide to Latent Dirichlet Allocation. https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158. Ημερομηνία πρόσβασης: 23/8/2018.

[31] Peng Junhui, Wei Wang, Ye qing Yu, Han lin Gu και Xuhui Huang. Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chinese Journal of Chemical Physics*, 31:404–420, 2018.

[32] K-Means Clustering – What it is and How it Works. http://www.learnbymarketing.com/methods/k-means-clustering/.

[33] Diederik P. Kingma και Jimmy Ba. Adam: A Method for Stochastic Optimization. Στο *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[34] Adrian Letchford, Helen Susannah Moat και Tobias Preis. The advantage of short paper titles. *Royal Society Open Science*, 2:150266, 2015.

[35] STEFANIA LOREDANA NITA. MACHINE LEARNING TECHNIQUES USED IN BIG DATA. *Scientific Bulletin of Naval Academy*, 19, 2016.

[36] Machine Learning. https://courses.edx.org/asset-v1:ColumbiaX+CSMM.101x+1T2017+type@asset+block@AI_edx_ml_5.1intro.pdf. Ημερομηνία πρόσβασης: 2019.

[37] Machine Learning: Bias VS. Variance. https://becominghuman.ai/machine-learning-bias-vs-variance-641f924e6c57. Ημερομηνία πρόσβασης: 11/10/2018.

[38] T. Soni Madhulatha. An Overview on Clustering Methods. *CoRR*, abs/1205.1117, 2012.

[39] Measuring a journal's impact. https://www.elsevier.com/authors/journal-authors/measuring-a-journals-impact.

[40] John Mingers και Loet Leydesdorff. A Review of Theory and Practice in Scientometrics. *European Journal of Operational Research*, 246(1):1–19, 2015.

[41] Multilayer Perceptron. https://deepai.org/machine-learning-glossary-and-terms/multilayer-perceptron.

[42] Natural Language Processing vs. Machine Learning vs. Deep Learning. https://rutumulkar.com/blog/2016/NLP-ML. Ημερομηνία πρόσβασης: 8/8/2016.

[43] NLP 101: Topic Modeling for Humans—Part #1. https://hackernoon.com/nlp-101-topic-modeling-for-humans-part-1-a030e8155584. Ημερομηνία πρόσβασης: 18/4/2018.

[44] Derek O'Callaghan, Derek Greene, Joe Carthy και Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Syst. Appl.*, 42(13):5645–5657, 2015.

[45] Enrique Orduna-Malea, Alberto Martín-Martín και Emilio Delgado López-Cózar. The next bibliometrics: ALMetrics (Author Level Metrics) and the multiple faces of author impact. *El Profesional de la Información*, 25:485, 2016.

[46] Robert Prohaska, Arnaud Konan, Kenneth Kelly και Michael Lammert. Heavy-Duty Vehicle Port Drayage Drive Cycle Characterization and Development. τόμος 9, 2016.

[47] Bapuji Rao και Mitra Anirban. A New Approach for Detection of Common Communities in a Social Network using Graph Mining Techniques. 2014.

[48] Bapuji Rao και Brojo Kishore Mishra. An Approach to Clustering of Text Documents Using Graph Mining Techniques. *IJRSDA*, 4(1):38–55, 2017.

[49] Scholarly Research Impact Metrics. https://rushu.libguides.com/metrics.

[50] Scientometrics of Scientometrics: Mapping Historical Footprint and Emerging Technologies in Scientometrics. https://www.intechopen.com/books/scientometrics/scientometrics-of-scientometrics-mapping-historical-footprint-and-emerging-technologies-in-scientome. Ημερομηνία πρόσβασης: 18/7/2018.

[51] Some thoughts on Journal Impact Factor. http://ivory.idyll.org/blog/2014-on-impact-factors.html.

[52] Iman Tahamtan, Askar Safipour Afshar και Khadijeh Ahamdzadeh. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107(3):1195–1225, 2016.

[53] Text Classification. A Comprehensive Guide to Classifying Text with Machine Learning. https://monkeylearn.com/text-classification/.

[54] The 5 Clustering Algorithms Data Scientists Need to Know. https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68. Ημερομηνία πρόσβασης: 5/2/2018.

[55] The Joy of Topic Modeling. http://mcburton.net/blog/joy-of-tm/. Ημερομηνία πρόσβασης: 21/5/2013.

[56] The mostly complete chart of Neural Networks, explained. https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464. Ημερομηνία πρόσβασης: 4/8/2017.

[57] Time Series for Dummies – The 3 Step Process. https://www.kdnuggets.com/2018/03/time-series-dummies-3-step-process.html.

[58] Types of Machine Learning Algorithms You Should Know. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861. Ημερομηνία πρόσβασης: 15/6/2017.

[59] Understanding Research Metrics: Journal-Level, Article-Level and Author-Level. https://www.enago.com/academy/what-are-different-research-metrics/. Ημερομηνία πρόσβασης: 24/9/2018.

[60] Jerome K. Vanclay. Factors affecting citation rates in environmental science. *J. Informetrics*, 7(2):265–271, 2013.

[61] Ulrikevon Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[62] What is Deep Learning? https://machinelearningmastery.com/what-is-deep-learning/. Ημερομηνία πρόσβασης: 16/8/2016.

[63] What is the difference between deep learning and usual machine learning? https://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning.

[64] What is underfitting and overfitting in machine learning and how to deal with it. https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. Ημερομηνία πρόσβασης: 11/3/2018.

[65] Why is the ReLU function not differentiable at x=0? https://sebastianraschka.com/faq/docs/relu-derivative.html.

[66] Why Use Metrics? Types of Journal Measures. https://www.enago.com/academy/eigenfactor-vs-impact-factor/.

[67] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M. Chu και Hongyuan Zha. On Modeling and Predicting Individual Paper Citation Count over Time. Στο *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, σελίδες 2676–2682, 2016.

[68] Siluo Yang. Are Scientometrics, Informetrics, and Bibliometrics different? Στο *The 16th International Conference on Scientometrics & Informetrics*, 2017.

[69] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan και Xiaoming Li. Citation count prediction: learning to estimate future citations for literature. Στο *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, σελίδες 1247–1252, 2011.

[70] ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ. `http://www.tmth.gr/home/59-applications/851-techniti-noymosyni`. Ημερομηνία πρόσβασης: 11/11/2013.

[71] Τεχνητή νοημοσύνη (AI). Τι είναι και γιατί είναι σημαντική. `https://www.sas.com/el_gr/insights/analytics/what-is-artificial-intelligence.html`.

[72] Υπολογιστική Νοημοσύνη. `https://eclass.pat.teiwest.gr/eclass/modules/document/file.php/589144/lesson2.pdf`.

Συντομογραφίες

LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
pLSI	Probabilistic Latent Semantic Indexing
TF-IDF	Term Frequency–Inverse Document Frequency
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DIC	Deviance Information Criterion
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
SSE	Sum of Squared Error
RMSE	Root Mean of Squared Error
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
RMSProp	Root Mean Square Propagation
API	Application Programming Interface
doc	document
seq2seq	sequence to sequence
lrate	learning rate
id	idem
T.N	Τεχνητή Νοημοσύνη
M.M	Μηχανική Μάθηση
B.M	Βαθιά Μάθηση
Ε.Φ.Γ	Επεξεργασία Φυσικής Γλώσσας
Σχ.	Σχήμα
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής

Ορολογία - Γλωσσάρι

Ελληνικός όρος	Αγγλικός όρος
Βαθιά Μάθηση	Deep Learning
Βαθιά Νευρωνικά Δίκτυα	Deep Neural Networks
Μηχανική Μάθηση	Machine Learning
απόκτηση δεδομένων	data acquisition
ανάπτυξη μοντέλου	deploy model
κρυμμένα στρώματα	hidden layers
χαρακτηριστικά	features
Εποπτευόμενη Μάθηση	Supervised Learning
Μη Εποπτευόμενη Μάθηση	Unsupervised Learning
Ενισχυμένη Μάθηση	Reinforcement Learning
Ημι-Εποπτευόμενη Μάθηση	Semi-Supervised Learning
Αυτοελεγχόμενη Μάθηση	Self-Supervised Learning
οπισθοδρόμηση	regression
ταξινόμηση	classification
συσταδοποίηση/ομαδοποίηση	clustering
άθροισμα των τετραγωνικών σφαλμάτων	sum of squared errors
ριζικό του μέσου αθροίσματος των τετραγωνικών σφαλμάτων	root mean of squared errors
Μείωση των διαστάσεων	Dimensionality reduction
Βιβλιομετρία	Bibliometrics
γνώρισμα	attribute
είσοδος	input
έξοδος	output
Δεδομένα Εκπαίδευσης	Train Data
Δεδομένα Επικύρωσης	Validation Data
Δεδομένα Δοκιμής	Test Data
υπερ-παράμετροι	hyper-parameters
Μοντέλο Πρόβλεψης	Predictive Model
επανάληψη	iteration
αναφορά/παραπομπή	citation
μετρικό	metric
παράγοντας αντίκτυπου	impact factor

δείκτης	index
αναγνωριστικό	idem
ετικέτα	label
Ταξινόμηση Κειμένου	Text Classification
μοντελοποίηση θέματος	topic modeling
Επεξεργασία Φυσικής Γλώσσας	Natural Language Processing
lemmatisation	λημματοποίηση
data preprocessing	προεπεξεργασία κειμένου
Document-Term Matrix	Πίνακας Εγγράφου-Όρου
κρυμμένα θέματα	latent topics
κατανομή	distribution
μίγμα	mixture
παράμετρος	parameter
όρος	term
πτώση/φθορά μάθησης	learning decay
προεξέχων	salient
λογαριθμική πιθανότητα	logarithmic likelihood
αντιστάθμιση	offset
σύγχυση	perplexity
τμηματικός	partitional
ιεραρχικός	hierarchical
συγκεντρωτικός	agglomerative
διαιρετικός	divisive
πυκνότητα	density
φασματικός	spectral
αρχικά σημεία	original points
ακραία περίπτωση	outlier
σύγκλιση	convergence
αγκώνας	elbow
ποσοστό	percentage
διακύμανση	variance
Αποσύνθεση Μοναδικής Τιμής	Singular Value Decomposition
διασταυρούμενη επικύρωση	cross validation
μεροληψία	bias
Χρονολογικές Σειρές	Time Series
Τάση	Trend
Εποχιακή Συνιστώσα	Seasonal Component
Τυχαία Συνιστώσα	Random Component
υπερ-προσαρμογή	over-fitting
υπο-προσαρμογή	under-fitting
παλινδρόμηση	regression

καθαρή είσοδος	net input
συνάρτηση μεταφοράς	transfer function
κανόνας της αλυσίδας	chain rule
Ανάστροφη Διάδοση Σφάλματος	Error Back Propagation
Ακτινική Λειτουργία Βάσης	Radial Basis Function
ρυθμός μάθησης	learning rate
ακολουθία	sequence
συνάρτηση ενεργοποίησης	activation function
όριο	threshold
Βηματική συνάρτηση	Step function
Γραμμική συνάρτηση	Linear function
Σιγμοειδής συνάρτηση	Sigmoid function
Διορθωμένη Γραμμική Μονάδα	Rectified Linear Unit
Συνελικτικά Νευρωνικά Δίκτυα	Convolution Neural Networks
Επαναλαμβανόμενο Νευρωνικό Δίκτυο	Recurrent Neural Network
κωδικοποιητής	encoder
αποκωδικοποιητής	decoder
πυκνό στρώμα	dense layer
κατανεμημένου χρόνου στρώμα	time distributed layer
στοχαστική κάθοδος κλίσης	stochastic gradient descent
αριθμός περιόδων χρόνου	epochs
μέγεθος παρτίδας	batch size
Διασύνδεση Προγραμματισμού Εφαρμογών	Application Programming Interface