Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών &
Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών

Διπλωματική Εργασία

Ανάπτυξη συστήματος ιεραρχικής ομαδοποίησης και διαχείρισης κειμένων για αποκεντρωμένα συστήματα ερωτοαπαντήσεων ορισμένου θέματος

Εκπόνηση:
Φώλας Δεμίρης Δημήτριος
ΑΕΜ: 9415

Επίβλεψη:
Καθ. Συμεωνίδης Ανδρέας
Υπ. Δρ. Μάλαμας Νικόλαος

Θεσσαλονίκη, Οκτώβριος 2023

The secret of getting ahead is getting started.
— Mark Twain

Ευχαριστιες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ.Ανδρέα Συμεωνίδη, για την εμπιστοσύνη που μου έδειξε σε όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, καθώς και για την καθοδήγησή του σε κρίσιμα σημεία αυτής.

Θα ήθελα επίσης να απευθύνω τις θερμές ευχαριστίες μου στον υποψήφιο διδάκτορα του τμήματος και του εργαστηρίου ISSEL, Νικόλαο Μάλαμα, για την συνεχή καθοδήγηση και επίβλεψη, τις πολύτιμες συμβουλές, την εύρυθμη συνεργασία και άμεση επικοινωνία του. Η εμπιστοσύνη των επιβλεπόντων κατά την εκπόνηση της παρούσας εργασίας αποτέλεσε θεμελειώδη παράγοντα και κινητήρια δύναμη.

Τέλος, ευχαριστώ την οικογένειά μου και τους φίλους μου για την κατανόηση και συνεισφορά τους σε προσωπικό επίπεδο.

Περίληψη

Η ανάκτηση πληροφορίας ανέκαθεν ήταν μια πλέον σηματνική πτυχή κάθε διεργασίας και δεδομένης της ταχείας αύξησης της απαίτησης γρήγορης και εύστοχης παροχής και ανάκτησης πληροφορίων, δεν είναι τίποτα παρά φυσικό να γίνεται συλλογική προσπάθεια προς την βελτιστοιποίηση αυτής της διαδικασίας με οποιαδήποτε μέσα είναι διαθέσιμα, όπως η Τεχνητη Νοημοσύνη. Με αυτό τον τρόπο θα μπορεί ένας υπολογιστής να "εκαπαιδευτεί" και να βοηθάει στο έργο αυτό, αντί να είναι μονάχα ένα εργαλείο για μαθηματικά πιθανοτήτων και στατιστική.

Σε μία εποχή όπου τα πάντα είναι καθοδηγούμενα από την πληροφορία και τα δεδομένα, η ανάγκη για δομημμένα δεδομένα και ορθή ανάκτηση πληροφορίας είναι τουλάχιστον επιτακτική. Η δομή και η οργάνωση στα δεδομένα διευκολύνει την λήψη αποφάσεων και μέσω αυτού επιβεβαιώνεται πάντα η σημασία και η συνεισφορά της τεχνητης νοημοσύνης και των μοντέλων μηχανικη μάθησης.

Η εφαρμογή και η υλοποίηση μεθόδων και τενικών βαθιάς μάθησης μπορεί σταδιακά μπορεί να βοηθήσει στην απαλλαγή μας από την εξάρτηση από λέξεις κλειδιά και να οδεύσουμε προς την διδασκαλία της σημασιολογικής κατανόησης της φυσικής γλώσσας από τους υπολογιστές. Σε αυτό μπορεί να συνεισφέρει εν μέρει από ένα πλήρως αυτόνομο σύστημα ικανό να οργανώνει, να διαχειρίζεται και να να ταξινομεί έγγραφα σημασιολογικά, με ελάχιστη εκπαίδευση. Η αξιοποίηση των δυνατοτήτων και της πολυχρηστικότητας της μάθησης και ταξινόμησης μηδενικών και λίγων βολών, καθώς και των σύγχρονων τεχνικών θεματικής μοντελοποίησης, μπορεί κανείς να αναπτύξει μια εφαρμογή που ως είσοδο λαμβάνει ακατέργαστα μη επισημειωμένα ή επεξεργασμένα δεδομένα και να επιστρέφει μια πλήρως λειτουργική εφαρμογή ερωτοαπαντήσεων.

Στην προσπάθεια μεγιστοποίσης της πολυχρηστικότητας του εν λόγω συστήματος, η παρούσα διπλωματική εργασία ερευνεί και αξιολογεί την βιωσιμότητα ενός συστήματος τέτοιας φύσεως. Δεδομένου ότι ο τελικός σκοπός είναι η δομημένη πληροφορία και η αποτελεσματική ανάκτηση της, θα εξεταστεί η υπόθεση της αποκεντρωμένης προσέγγισης, καθώς μειώνονται σημαντικά οι απαιτήσεις υπολογιστικής ισχύος και αποθηκευτικού χώρου.

Μέσω δοκιμών και πειραμάτων τα παραγώμενα δεδομένα φαίνεται να υποστηρίζουν την υπόθεση υπέρ ενός τέτοιου συστήματος, και δυνητικά με διάφορα πλεονεκτήματα υπέρ ενός αντίστοιχου αλλά ενιαίου συστήματος.

Φώλας Δεμίρης Δημήτριος
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Νοέμβριος 2023

A System for Semantic Hierarchical Clustering and Management of Text Documents for Automated Domain-Specific Decentralized Question Answering Systems

Abstract

Information retrieval has always been a fundamental part of almost any task, therefore as the world steadily increases its reliance on quick and accurate information retrieval, it is only natural to focus on finding ways, with which one could leverage the power of artificial intelligence in order to improve the question answering capabilities of a computer system, rather than to stay reliant on more mature, merely probabilistic and statistical methods.

In an era where data drives innovation and decision-making, the importance of structured and organized data cannot be overstated. Structured data facilitates efficient information retrieval, supports data-driven decision-making processes, and empowers artificial intelligence and machine learning models.

The implementation of deep learning methods and techniques can gradually aid the procedure to relieve us of our dependence on keywords and shift towards teaching a computer to understand the semantic capacity of text documents and queries and use that as a means of a more refined method of information retrieval. The former can be aided in part by a fully autonomous system capable of organizing, managing and classifying documents semantically with minimal training. Harnessing the capabilities and profound versatility of zero and few-shot learning and classification, as well as modern topic modelling techniques, one can develop an application that as an input receives raw unlabelled and unprocessed data and returns a fully functioning information retrieval and question-answering system.

In an effort to be versatile throughout its premise, within this research, the feasibility and profitability of such a system being decentralized and domain-specific will be challenged and assessed. As the end goal is structured data and efficient information retrieval, the possibility of a system functioning within nodes that are separate from each other but communicate implies its dependency on smaller less powerful and easier-to-maintain systems, with less storage and computational requirements and therefore is a promising hypothesis.

Through an assortment of benchmarks and tests the data supports the hypothesis that such a system if planned and executed according to the problem at hand, could be a better and more versatile solution.

Folas Demiris Dimitrios
Electrical & Computer Engineering Department,
Aristotle University of Thessaloniki, Greece

November 2023

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος πινάκων

Ακρωνύμια Εγγράφου

Παρακάτω παρατίθενται ορισμένα από τα πιο συχνά χρησιμοποιούμενα ακρωνύμια της παρούσας διπλωματικής εργασίας:

AI	$\longrightarrow$	Artificial Intelligence
ANN	$\longrightarrow$	Artificial Neural Netwrkk
BERT	$\longrightarrow$	Bidirectional Encoder Representations from Transformers
CNN	$\longrightarrow$	Convolutional Neural Networks
DL	$\longrightarrow$	Deep Learning
DNN	$\longrightarrow$	Deep Neural Networks
HDBSCAN	$\longrightarrow$	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA	$\longrightarrow$	Latent Dirichlet Allocation
LLM	$\longrightarrow$	Large Language Model
LSA	$\longrightarrow$	Latent Sematntic Analysis
LSTM	$\longrightarrow$	Long Short Term Memory
ML	$\longrightarrow$	Machine Learning
NLI	$\longrightarrow$	Natural Language Inference
NLP	$\longrightarrow$	Natural Language Programmming
NLU	$\longrightarrow$	Natural Language Understanding
NNMF	$\longrightarrow$	Non Negative Matrix Factorization
QA	$\longrightarrow$	Question Answering
RNN	$\longrightarrow$	Recurrent Neural Network
TF	$\longrightarrow$	Transformer
TF-IDF	$\longrightarrow$	Term Frequency Inverse Document Frequency

1 Εισαγωγή

1.1 Περιγραφη του Προβληματος

Η τρομακτική ανάπτυξη του διαδικτύου και της τεχνολογίας και η ενσωματωσή του στην καθημερινότητά μας μπορεί να διευκολύνουν την πλειοψηφία των δραστηριοτήτων ή και των αναγκών μας, αλλά παράλληλα καθιστούν επιτακτική ανάγκη την ανάπτυξη συστημάτων και μηχανισμών που να μπορούν να συμβαδίζουν με την τεράσια ροή ανεπεξέργαστης πληροφορίας. Είναι πιο σημαντικό από ποτέ να βελτιστοποιηθεί ο τρόπος με τον οποίο διαχειρίζονται, ομαδοποιούνται, ταξινομούνται και ανακτώνται οι πληροφορίες, καθώς είναι ο μόνος τρόπος να υπάρχει τάξη και να μπορεί η πολύτιμη πληροφορία να αξιοποιηθεί στην πληρότητά της. Ο όγκος των δεδομένων που παράγεται πλέον είναι τόσο μεγάλος που καθίσταται μηδενική η πιθανότητα να μπορεί να οργανωθεί και να γίνει διαχειρίσιμο το σύνολό τους χειροκίνητα. Απαιτείται επομένως ένα σύστημα που να μπορεί αποτελεσματικά να οργανώνει και να ταξινομεί δεδομένα, ώστε ύστερα αυτά να μπορούν να ανακτηθουν με ευκολία εφόσον ζητηθούν. Ταυτόχρονα αυτό θα πρέπει να γίνεται με τρόπο τέτοιο, ούτως ώστε να μην σπαταλούνται υπολογιστικοί ή αποθηκευτικοί πόροι. Η ανάκτηση πληροφορίας μέχρι πρότινος ήταν πρόβλημα αναζήτησης λέξεων κλειδιά, αλλά είναι πλέον εύκολο για έναν υπολογιστή να καταλαβαίνει το νόημα πίσω από τις λέξεις, κάτι το οποίο μπορούμε να εκμεταλλευτούμε. Παρ' όλα αυτά όμως λόγω του όγκου των δεδομένων σύντομα οι παραδοσιακοί τρόποι ανάκτησης δεδομένων είτε βάσει λέξεων κλειδιά ή με μοντέλα βαθιάς μάθησης και τεχνητής νοημοσύνης θα σταματήσουν να είναι αποδοτικοί, καθώς θα απαιτούν τεράστια υπολογιστική δύναμη, χρόνο για την αναδρομή σε αμέτρητα δεδομένα και συντριπτικά μεγάλο χώρο αποθήκευσης. Φαίνεται συνεπώς ότι ενδεχομένως να υπάρχει η ευκαιρία να προοδεύσει μια προσέγγιση διαφορετική από τη φιλοσοφία των τεραστίων γλωσσικών μοντέλων και δεδομένων εκπαίδευσης, μια προσέγγιση πιο αυτόνομη που να μπορεί να ταξινομεί κείμενα, να τα οργανώνει και να τα διαχειρίζεται με τρόπο τέτοιο ώστε να μην είναι όλα συγκεντρωμένα μαζί, αλλά να δημιουργούν διακριτές

μεταξύ τους ομάδες θεματολογίας ή κάποιου άλλου μετρήσιμου μεγέθους και να επικοινωνούν μεταξύ τους μόνο αν αυτό χρειαστεί.

1.2 Σκοπος - Συνεισφορα της Διπλωματικης Εργασιας

Ο στόχος της παρούσας διπλωματικής εργασίας είναι ακριβώς αυτό που περιγράφεται παραπάνω. Να αναπτυχθεί ένα σύστημα ως proof-of-concept για αποκεντρωμένη προσέγγιση στην δημιουργία συστημάτων διαχείρησης και ανάκτησης πληροφορίας σε μορφή φυσικής γλώσσας. Ένα σύστημα αυτόνομο που να μπορεί να δέχεται ως είσοδο ανεπεξέργαστα κείμενα ή έγγραφα και να τα αναλύει, να τα ταξινομεί και να τα ομαδοποιεί με τρόπο, τέτοιο που δεν χρειάζεται να βρίσκονται τα πάντα μαζί για να λειτουργεί. Αυτό αφενώς μεν μειώνει τις απαιτήσεις για χώρο αποθήκευσης, σε περίπτωση πολύ μεγάλων συνόλων δεδομένων, αφετέρου δε μειώνει αισθητά την απαιτούμενη υπολογιστική ισχύ για να λειτουργεί εύρυθμα. Με την αξιοποίηση τεχνικών ταξινόμηση μηδενικλής βολής δύνανται να παρακαμφθούν κοστοβόρες διαδικασίες εκπαίδευσης ή διανυσματικής αναπαράστασης. Ένα τέτοιο σύστημα επίσης αν λειτουργεί με την προσδοκούμενη ακρίβεια θα σημαίνει ότι σε μεγάλο βαθμό τα κείμενα θα μπορούν να οργανώνονται αυτόματα χώρις επιπλέουσα χειροκίνητη παρέμβαση και θα οδεύουμε σε ένα concept διαφάνειας δεδομένων. Επιπρόσθετα εφόσον ένα μεγάλο σύστημα διασπάται ή αξιοποιείται εν μέρει, αυτό σημαίνει ότι ενδεχομένως να υπάρχουν περισσότερες ευκαιρίες παραμετροποίησης στα εκάστοτε δεδομένα και απαιτήσεις χωρίς να απαιτεί αυτό θεμελειώδεις αλλαγές. Τέλος ένα υπολογιστικό σύστημα αποκεντρωμένο, εν γένει φέρει πλεονεκτήματα ασφαλείας από καταστροφές ή βλάβες υλικού, που αντί να θέσουν εκτός λειτουργίας ολόκληρο το σύστημα, παύει να λειτουργεί μόνο το στοιχείο που παρουσίασε την βλάβη. Συνεπώς στόχος είναι η αξιολόγηση τόσο της βιωσιμότητας ενός συστήματος αυτής της φιλοσοφίας όσο και των διαφορών του με ένα πιο απλό ενιαίο σύστημα.

1.3 Διαρθρωση της Αναφορας

Η διάρθρωση της παρούσας διπλωματικής εργασίας είναι η εξής:

- **Κεφάλαιο 2**: Γίνεται μια γενικη επισκόπηση της ευρύτερης ερευνητικής περιοχής και αναδρομή σε προηγούμενες μελέτες.
- **Κεφάλαιο 3**: Γίνεται η παρουσίαση του θεωρητικού υπόβραθρου, πάνω στο οποίο βασίζονται οι προκείμενες μεθοδολογίες και προσεγγίσεις.
- **Κεφάλαιο 4**: Παρουσιάζεται και αναλύεται η μεθοδολογία που υλοποιείται για την δημιουργία και την λειτουργία του συστήματος.
- **Κεφάλαιο 5**: Γίνεται παρουσίαση και αξιολόγηση των δοκιμών του συστήματος και των αποτελεσμάτων αυτών.

- **Κεφάλαιο 6**: Παρουσιάζονται τα τελικά αποτελέσματα και συμπεράσματα της εργασίας και θέματα για μελλοντική μελέτη ή δυνητικές επεκτάσεις.

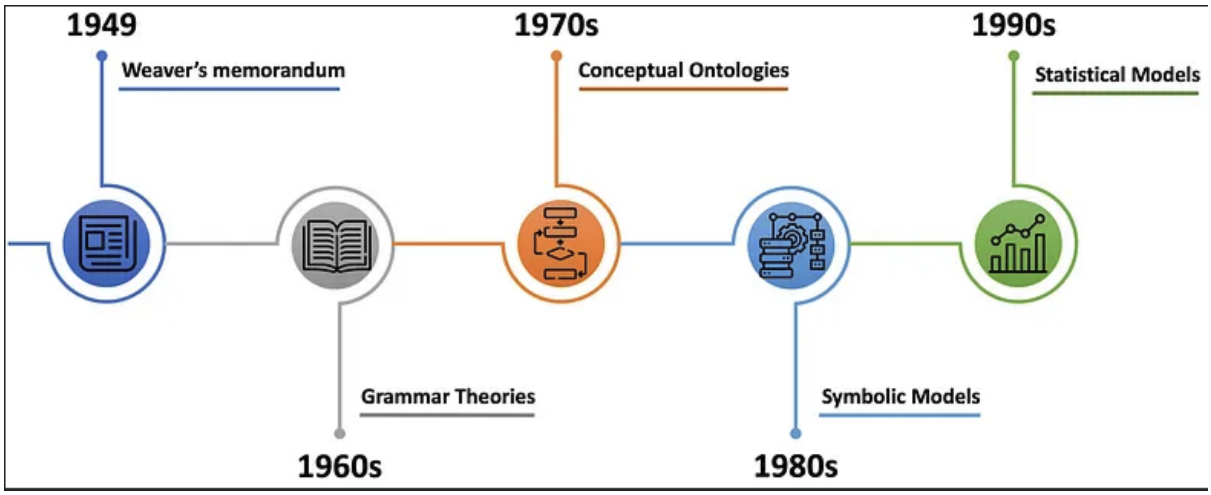
2
Επισκόπηση Ερευνητικής Περιοχής

2.1 Επεξεργασια Φυσικη Γλωσσας - NLP

Η επεξεργασία φυσικής γλώσσας (NLP) είναι ένα θεωρητικά αιτιολογημένο φάσμα υπολογιστικών τεχνικών για την ανάλυση και αναπαράσταση κειμένων που συμβαίνουν στη φύση σε ένα ή περισσότερα επίπεδα γλωσσικής ανάλυσης (Liddy, 2001)[2]. Σκοπός αυτών των τεχνικών είναι να επιτευχθεί γλωσσική επεξεργασία που μοιάζει με την ανθρώπινη για μια σειρά εργασιών ή εφαρμογών. Παρόλο που έχει αποκτήσει τεράστιο ενδιαφέρον τα τελευταία χρόνια, η έρευνα στον τομέα της NLP ξεκίνησε πριν από αρκετές δεκαετίες και χρονολογείται από τα τέλη της δεκαετίας του 1940. Είναι κοινή παραδοχή ότι οι πρώτες ιδέες για τις πρώτες εφαρμογές βασισμένες σε υπολογιστή σχετικές με την φυσική γλώσσα ξεκίνησαν με το υπόμνημα Weaver - Weaver's Memorandum (Shannon and Weaver, 1949)[3]. Αποτέλεσε έμπνευση για επακόλουθες έρευνες όπως το πείραμα Georgetown - Georgetown Experiment, το 1954, κατά το οποίο επιτυχώς μεταφράστηκαν με υπολογιστή περισσότερα από 60 ρώσικα κείμενα στα αγγλικά, το οποίο, όμως, χρησιμοποιούσε εξειδικευμένα σύνολα κανόνων και απέτυχε να μπορέσει να χρησιμοποιηθεί για εύτερες εφαρμογές μηχανικής μετάφρασης - machine translation (MT).

Στη συνέχεια, οι ερευνητές συνειδητοποίησαν σταδιακά ότι το έργο ήταν πολύ πιο δύσκολο από ό,τι περίμεναν και ότι χρειάζονταν μια πιο επαρκή θεωρία της γλώσσας. Χρειάστηκε να φτάσει το 1957 για να εισαχθεί η ιδέα της γενεσιουργού γραμματικής (Chomsky, 1957)[4], ένα σύστημα συντακτικών δομών βασισμένο σε κανόνες, το οποίο έφερε την εικόνα για το πώς η κυρίαρχη γλωσσολογία θα μπορούσε να βοηθήσει τη μηχανική μετάφραση. Λόγω της ανάπτυξης της συντακτικής θεωρίας της γλώσσας και των αλγορίθμων ανάλυσης, η δεκαετία του 1950 κατακλύστηκε από υπερβολικό ενθουσιασμό. Ο κόσμος πίστευε ότι τα πλήρως αυτόματα συστήματα μετάφρασης υψηλής ποιότητας θα ήταν σε θέση να παράγουν αποτελέσματα που δεν θα μπορούσαν να διαφέρουν από εκείνα των ανθρώπινων μεταφρα-

στών και ότι τέτοια συστήματα θα λειτουργούσαν μέσα σε λίγα χρόνια. Δεδομένων των τότε διαθέσιμων γλωσσολογικών γνώσεων και των συστημάτων υπολογιστών, η σκέψη αυτή ήταν εντελώς μη ρεαλιστική. Αφότου είχε περάσει παραπάνω από μια δεκαετία ερευνών εξακολουθούσε να μην μπορεί να συγκιθεί η μετάφραση των υπολογιστών με αυτή ενός ανθρώπου και φάνηκε ότι η πορεία της έρευνας απείχε πολύ από την προσεχή ευρεία εφαρμοφή της, σε σημείο που επιστημονικές επιτροπές παρότρυναν την διακοπή χρηματοδότησης ερευνών, Automatic Language Processing Advisory Committe - ALPAC, The (in)famous report, 1966[5], το οποίο κατέληξε στο συμπέρασμα ότι η MT δεν ήταν άμεσα εφικτή και συνέστησε στην ερευνητική κοινότητα να σταματήσει τη χρηματοδότησή της. Αυτό είχε ως αποτέλεσμα την ουσιαστική επιβράδυνση όχι μόνο της έρευνας MT, αλλά και των περισσότερων εργασιών σε άλλες εφαρμογές του NLP. Ύστερες έρευνες βασίστηκαν σε μελέτες για τον τρόπο με τον οποίο να μπορεί να αποδοθεί νόημα από την φυσική γλώσσα στους υπολογιστές και την δημιουργία κανόνων γραμματικής. Εκείνη την εποχή επίσης πρωτοεμφανήστηκαν τα πρώτα συστήματα συζήτησης, όπως μεταξύ άλλων και η ELIZA, ένα σύστημα συζήτησης - chatbot προσπάθησε να προσωμοιώσει τεχνικές ψυχιατρικής.



Σχήμα 2.1: Στάδια της NLP πριν την Βαθιά Μάθηση Πηγή: *https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce*

Η δεκαετία του 1970 έφερε νέες ιδέες στο NLP, όπως η δημιουργία εννοιολογικών οντολογιών που δομούσαν τις πληροφορίες του πραγματικού κόσμου σε δεδομένα κατανοητά από τον υπολογιστή. Κατά την δεκαετία του 1990, κυριάρχησαν τα στατιστικά μοντέλα στην NLP και κατάφεραν να αντικαταστήσουν την πλειοψηφία των εκάστοτε "χειρόγραφων" περίπλοκων κανόνων, καθώς ήταν υποβοηθούμενα από την συνεχώς αυξανόμενη υπολογιστική ισχύ και αλγόριθμους μηχανικής μάθησης, όπως τα δένδρα απόφασης. Τα στατιστικά μοντέλα ξεπέρασαν το φράγμα πολυπλοκότητας των χειροκίνητα κωδικοποιημένων κανόνων, δημιουργώντας τους μέσω αυτόματης μάθησης, γεγονός που οδήγησε την έρευνα να εστιάζει όλο και περισσότερο σε αυτά τα μοντέλα. Εκείνη την εποχή, αυτά τα στατιστικά μοντέλα ήταν ικανά να λαμβάνουν ήπιες, πιθανολογικές αποφάσεις.

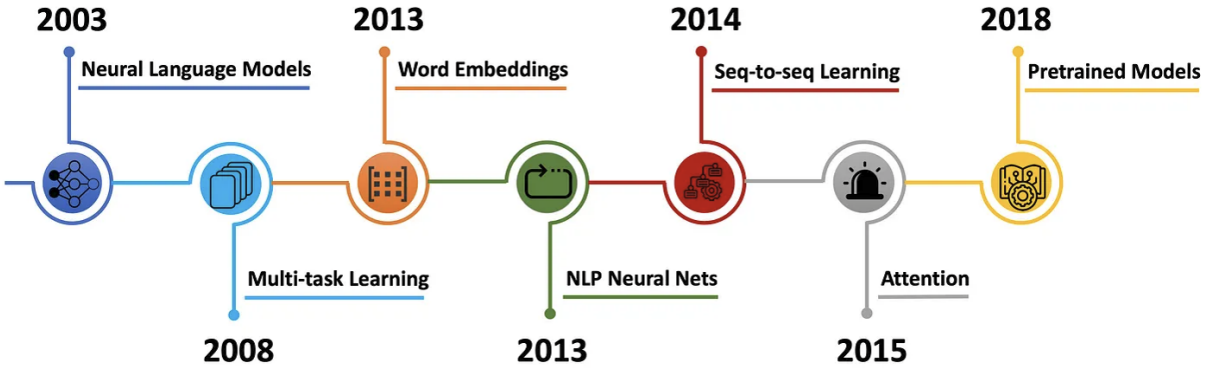
Το επόμενο άλμα διαδραματίζεται στην επόμενη δεκαετία, όπου πρωτοχρησιμοποιούνται τα νευρωνικά δίκτυα - artificial neural networks - ANN, με τα οποία οι λέξεις αναπαριστώνταν διανυσματικά - embedding, λαμβάνοντας υπόψη τις n προηγούμενες λέξεις. Το 2013 εμφανίζεται ένα από τα πιο δημοφιλή μοντέλα διανυσματικής αναπαράστασης, το Word2Vec, Mikolov et al[6], το οποίο επέτρεψε για πρώτη φορά την εκπαίδευση σε τεράστια σύνολα κειμένων - corpora, χάρη στην αποδοτική υλοποίηση του. Ακολούθησαν ιδέες και υλοποιήσεις όπως τα Long Short Term Memory Networks με Νευρωνικά Δίκτυα με Ανατροφοδότηση - Recurrent Neural Networks (RNNs) και η έννοια της προσοχής - attention μέχρι την εισαγωγή του Μετασχηματιστή - Transformer, με την αρχιτεκτονική του οποίου, μοντέλα αποδίσουν εξαιρετικά αποτελέσματα. Η τελευταία μεγάλη καινοτομία στον κόσμο του NLP είναι αναμφίβολα τα μεγάλα προ-εκπαιδευμένα γλωσσικά μοντέλα - pre-trained language models. Αν και προτάθηκαν για πρώτη φορά το 2015 (Dai και Le)[7], αργότερα αποδείχθηκε ότι προσφέρουν μεγάλη βελτίωση σε σχέση με τις σύγχρονες μεθόδους σε ένα ευρύ φάσμα εργασιών. Οι προ-εκπαιδευμένες διανυσματικές αναπαραστάσεις γλωσσικών μοντέλων μπορούν να χρησιμοποιηθούν ως χαρακτηριστικά σε ένα μοντέλο-στόχο - target model (Peters et al., 2018)[8], ή ένα προ-εκπαιδευμένο γλωσσικό μοντέλο - pretrained language model μπορεί να ρυθμιστεί λεπτομερώς σε δεδομένα εργασίας-στόχου - target-task (Devlin et al., 2018[9]- Howard and Ruder, 2018[10]- Radford et al., 2019[11]- Yang et al., 2019[12]), τα οποία έχουν δείξει ότι επιτρέπουν αποτελεσματική μάθηση με σημαντικά λιγότερα δεδομένα. Μεγάλο πλεονέκτημα των προ-εκπαιδευμένων γλωσσικών μοντέλων είναι η ικανότητά τους να μαθαίνουν γλωσσικές αναπαραστάσεις από τεράστια μη σεσημασμένα σύνολα κειμένων - unannotated text corpora, γεγόνος που οφελεί σημαντικά γλώσσες με έλλειψη ταξινομημένων και σεσημασμένων γλωσσικών δεδομένων.

Η Επεξεργασία Φυσικής Γλώσσας - NLP μπορεί να ταξινομηθεί σε δύο μέρη, δηλαδή στην Κατανόηση Φυσικής Γλώσσας - Natural Language Understanding (NLU) και στη Παραγωγή Φυσικής Γλώσσας - Natural Language Generation (NLG).

Η Κατανόηση Φυσικής Γλώσσας - NLU επιτρέπει στις μηχανές να κατανοούν τη φυσική γλώσσα και να την αναλύουν εξάγοντας έννοιες, οντότητες, συναισθήματα, λέξεις-κλειδιά κ.λ.π.. Χρησιμοποιείται σε εφαρμογές για την κατανόηση των αιτημάτων/μηνυμάτων που αναφέρει κανείς είτε προφορικά είτε γραπτά. Η γλωσσολογία - linguistics είναι η επιστήμη που περιλαμβάνει τη σημασία της γλώσσας, το γλωσσικό πλαίσιο και τις διάφορες μορφές της γλώσσας. Έτσι, είναι σημαντικό να κατανοήσουμε διάφορες σημαντικές ορολογίες του NLP και τα διάφορα επίπεδα του NLP. Η παραγωγή φυσικής γλώσσας - NLG είναι η διαδικασία παραγωγής φράσεων, προτάσεων και παραγράφων που έχουν νόημα από μια εσωτερική αναπαράσταση. Αποτελεί μέρος της Επεξεργασίας Φυσικής Γλώσσας - NLP και πραγματοποιείται σε τέσσερις φάσεις: προσδιορισμός των στόχων, σχεδιασμός του τρόπου με τον οποίο μπορούν να επιτευχθούν οι στόχοι αξιολογώντας την κατάσταση και τις διαθέσιμες επικοινωνιακές πηγές και υλοποίηση των σχεδίων ως κείμενο. Είναι το αντίθετο/αντίστροφο της Κατανόησης Φυσικής Γλώσσας - NLU.

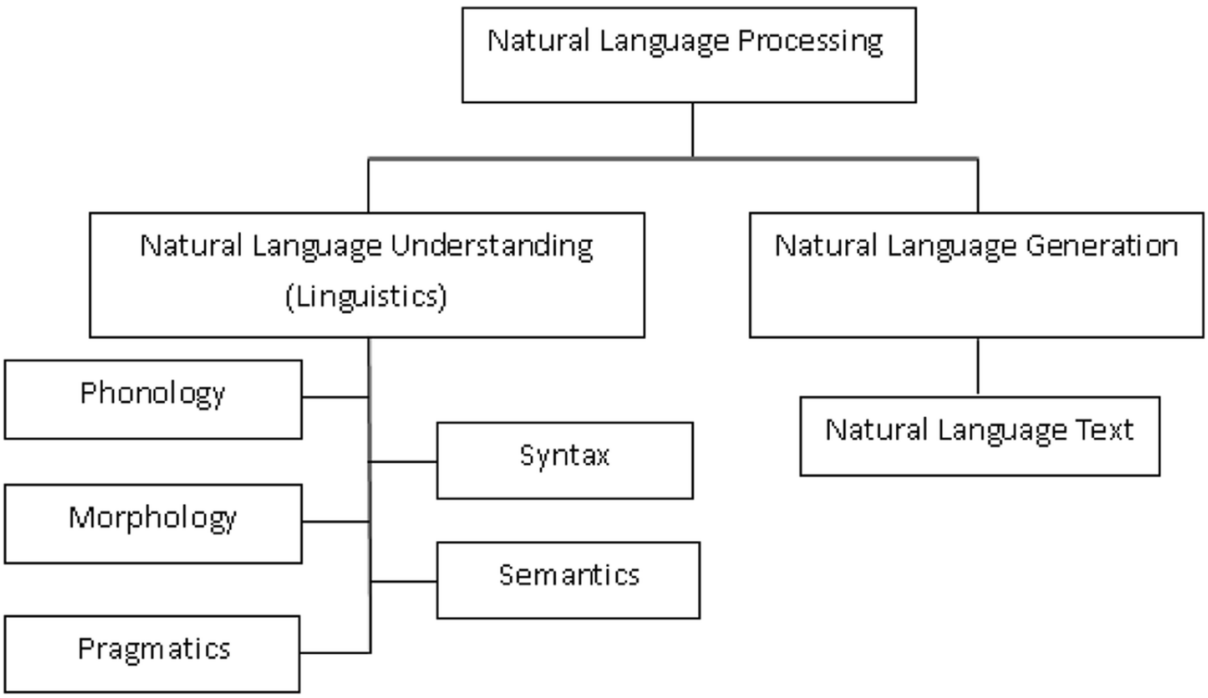
Η Επεξεργασία Φυσικής Γλώσσας - NLP μπορεί να εφαρμοστεί σε διάφορους το-

μείς, όπως η μηχανική μετάφραση - Machine Translation (MT), η ταξινόμηση κειμένου - text classification, η εξαγωγή και ανάκτηση πληροφοριών - data extraction and retrieval, η σύνοψη - summarization, η απάντηση ερωτήσεων - question answering (QA) κ.λ.π.. Η παρούσα εργασία, σαφώς, επικεντρώνεται κυρίως στην ταξινόμηση κειμένου - text classification και την απάντηση ερωτήσεων μέσω ανάκτησης πληροφοριών - Extractive Question Answering - QA.



Σχήμα 2.2: Στάδια της σύγχρονης NLP Πηγή: *https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37*

Πιο συγκεκριμένα θα συνδυαστούν δυο καινοτόμες προσεγγίσεις για την επίλυση του προβλήματος της ταξινόμησης κειμένων, η θεματική μοντελοποίηση μέσω μετασχηματιστών - transformer based topic modelling και η ταξινόμηση μηδενικής βολής - zero shot classification, η ύστερη δε αποτελεί αντικείμενο έλξης για πρόσφατες έρευνες δεδομένης της προσαρμοστικότητας και ευκολίας της τεχνικής αυτής.



Σχήμα 2.3: Διάρθρωση επιστήμης NLP

2.2 Θεματικη Μοντελοποιηση - Topic Modelling

Τα θεματικά μοντέλα έχουν εφαρμοστεί σε οτιδήποτε, από βιβλία μέχρι εφημερίδες και αναρτήσεις στα μέσα κοινωνικής δικτύωσης, σε μια προσπάθεια να εντοπιστούν τα πιο διαδεδομένα θέματα ενός σώματος κειμένων. Παρέχουμε μια εμπεριστατωμένη ανάλυση των μη επιβλεπόμενων θεματικών μοντέλων από την ίδρυσή τους μέχρι σήμερα. Εντοπίζουμε την προέλευση των διαφόρων τύπων σύγχρονων θεματικών μοντέλων, ξεκινώντας από τη δεκαετία του 1990.

Το θεματικό μοντέλο είναι ένα είδος πιθανοτικού παραγωγικού μοντέλου που έχει χρησιμοποιηθεί ευρέως στον τομέα της επιστήμης των υπολογιστών με ιδιαίτερη έμφαση στην εξόρυξη κειμένου και στην ανάκτηση πληροφοριών. Από τότε που προτάθηκε για πρώτη φορά αυτό το μοντέλο, έχει λάβει πολλά προσοχή και απέκτησε ευρύ ενδιαφέρον μεταξύ των ερευνητών σε πολλά ερευνητικά πεδία. Έτσι, το μέχρι στιγμής, εκτός από την εξόρυξη κειμένου, έχουν επίσης υπάρξει επιτυχημένες εφαρμογές στους τομείς της όρασης υπολογιστών (Fei-Fei and Perona 2005[13] - Luo et al. 2015[14]), της γενετικής πληθυσμών και κοινωνικά δίκτυα (Jiang et al. 2015)[15].

Η προέλευση ενός θεματικού μοντέλου είναι η λανθάνουσα σημασιολογική δεικτοδότηση - Latent Semantic Indexing (LSI) (Deerwester et al. 1990[16]) και έχει υπάρξει βάση της ανάπτυξης ενός θεματικού μοντέλου. Ωστόσο, το LSI δεν είναι ένα πιθανοτικό μοντέλο, συνεπώς, δεν είναι ένα αυθεντικό θεματικό μοντέλο. Με βάση την LSI, η Πιθανολογική Λανθάνουσα Σημασιολογική Ανάλυση - Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 2001) προτάθηκε από τον Hofmann και είναι ένα γνήσιο θεματικό μοντέλο. Δημοσιευμένο μετά την PLSA, η Λανθάνουσα Κατανομή Dirichlet - Latent Dirichlet Allocation (LDA) που προτάθηκε από τους Blei et al. (2003)[17] είναι ένα ακόμη πιο πλήρες πιθανοτικό παραγωγικό μοντέλο και αποτελεί την επέκταση της PLSA. Σήμερα, υπάρχει ένας αυξανόμενος αριθμός πιθανοτικών μοντέλων που βασίζονται στην LDA μέσω του συνδυασμού συγκεκριμένων εργασιών. Παρ' όλα αυτά, όλα τα προαναφερθέντα θεματικά μοντέλα εισήχθησαν αρχικά στην κοινότητα της ανάλυσης κειμένου για μη επιβλεπόμενη ανακάλυψη θεμάτων σε ένα σώμα εγγράφων.

Ο Top2Vec (Angelov, 2020[18]) είναι ένας συγκριτικά νέος αλγόριθμος που χρησιμοποιεί διανυσματική αναπαράσταση λέξεων. Δηλαδή, η διανυσματοποίηση των γλωσσικών δεδομένων καθιστά δυνατό τον εντοπισμό σημασιολογικά όμοιων λέξεων, προτάσεων ή και εγγράφων σε χωρική εγγύτητα (Egger, 2022a[19]). Για παράδειγμα, λέξεις όπως "στυλό" και "μολύβι" θα πρέπει να βρίσκονται πιο κοντά από λέξεις όπως "γραφείο" και "σκύλος".

Το BERTopic (Grootendorst, 2020)[9] βασίζεται στους μηχανισμούς του Top2Vec και ως εκ τούτου, είναι παρόμοιες από την άποψη της αλγοριθμικής δομής. Όπως υποδηλώνει το όνομα, αξιοποιούνται μοντέλα μετασχηματιστών BERT για την παραγωγή διανυσματικών αναπαραστάσεων του κειμένου και των εγγράφων - text and document embeddings, και το BERTopic παρέχει εξαγωγή ενσωμάτωσης εγγράφων, με ένα μοντέλο μετασχηματιστών προτάσεων για περισσότερες από 50

γλώσσες. Ομοίως, το BERTopic υποστηρίζει επίσης το UMAP για διάσταση και HDBSCAN ή k-Means για ομαδοποίηση εγγράφων. Μια ειδοποιός διαφορά μεταξύ του BERTopic και του Top2Vec είναι η εφαρμογή ενός αλγορίθμου αντίστροφης συχνότητας εγγράφων με βάση τις κλάσεις - class based term frequency inverse document frequency (c-TF-IDF), ο οποίος συγκρίνει τη σημασία των όρων εντός ενός συστάδα και δημιουργεί αναπαράσταση όρων (Sánchez-Franco and Rey-Moreno, 2022). Αυτό σημαίνει ότι όσο υψηλότερη είναι η τιμή ενός όρου, τόσο πιο αντιπροσωπευτικός είναι αυτός για το θέμα του.

2.3 Συστηματα Ερωτοαπαντησεων - QA Systems

Η ταχεία αύξηση της μαζικής αποθήκευσης πληροφοριών και η δημοτικότητα της χρήσης του διαδικτύου επιτρέπουν σε όλους τους χρήστες να αποθηκεύουν δεδομένα και να τα καταστήσουν διαθέσιμα στο κοινό. Ωστόσο, η εξερεύνηση αυτού του μεγάλου όγκου δεδομένων καθιστά την εύρεση πληροφοριών ένα πολύπλοκο και δαπανηρό έργο, τόσο από άποψη χρόνου, αλλά και από άποψη πόρων. Αυτή η δυσκολία έχει δώσει κίνητρο για την έρευνα και την ανάπτυξη νέων προσαρμοσμένων ερευνητικών εργαλείων, όπως τα συστήματα απάντησης ερωτήσεων.

Τα αρχικά στάδια συστημάτων ερωτοαπαντήσεων βασίζονταν σε ακριβή ταύτηση όρων ή λέξεις κλειδιά - keyword based search. Η αναζήτηση με λέξεις-κλειδιά υπάρχει εδώ και πολύ καιρό και λειτουργεί όπως το ευρετήριο στο τέλος ενός βιβλίου. Μια μηχανή αναζήτησης λέξεων-κλειδιών δημιουργεί ένα ευρετήριο όλων των λέξεων σε όλα τα έγγραφα και παρέχει αποτελέσματα με βάση απλούς αλγορίθμους αντιστοίχισης. Για να βελτιώσουν τη συνάφεια της αναζήτησης και την κατάταξη των αποτελεσμάτων, εισήχθησαν στατιστικά στοιχεία λέξεων, όπως το TF-IDF[20].

Η στατιστική κατάταξη ήταν χρήσιμη, αλλά όχι αρκετή. Υπήρχαν πάρα πολλές περιπτώσεις χρήσης όπου οι λέξεις δεν ταίριαζαν ακριβώς με το ερώτημα. Για παράδειγμα, όροι στον ενικό έναντι του πληθυντικού, κλίσεις ρημάτων (ενεστώτας έναντι αορίστου, παρατατικός, κ.λπ.), συγκολλητικές ή σύνθετες γλώσσες κ.λ.π.. Αυτό οδήγησε στην ανάπτυξη λειτουργιών επεξεργασίας φυσικής γλώσσας (NLP) για τη διαχείριση της πολυπλοκότητας των γλωσσών. Μια άλλη μέθοδος για την ανάπτυξη μιας καλύτερης, σημασιολογικής κατανόησης ενός ερωτήματος ήταν η χρήση οντολογιών και γραφημάτων γνώσης. Οι γράφοι γνώσης αναπαριστούν μια σχέση μεταξύ διαφορετικών στοιχείων - εννοιών, αντικειμένων, γεγονότων. Μια οντολογία ορίζει καθένα από τα στοιχεία και τις ιδιότητές τους. Αυτή συνεπώς η σημασιολογική προσέγγιση προσπάθησε να αναπαραστήσει διαφορετικές έννοιες και τις συνδέσεις μεταξύ τους.

Από το 2013[6] και μετά εντάσσεται σε αυτό τον κλάδο η θεωρία για την διανυσματική αναζήτηση - vector search. Η διανυσματική αναπαράσταση του κειμένου είναι πολύ παλιά. Οι θεωρητικές της ρίζες ανάγονται στη δεκαετία του 1950 και υπήρξαν αρκετές σημαντικές εξελίξεις κατά τη διάρκεια των δεκαετιών. Στην

απλούστερη μορφή της, η διανυσματική αναζήτηση, είναι ένας τρόπος για την εύρεση σχετικών αντικειμένων που έχουν παρόμοια χαρακτηριστικά. Η αντιστοίχιση επιτυγχάνεται με μοντέλα μηχανικής μάθησης που ανιχνεύουν σημασιολογικές σχέσεις μεταξύ αντικειμένων σε ένα ευρετήριο. Τα διανύσματα μπορεί να έχουν χιλιάδες διαστάσεις. Κύριο πρωτέρημα της διανυσαμτικής προσέγγισης είναι η διαχείρηση ερωτήσεων σε μορφή προτάσεων αντί για μεμονομένες λέξεις, στις οποίες περιπτώσεις ενδεχομένως η αναζήτηση με λέξεις κλειδιά να έχε καλύτερα αποτελέσματα.

3
Θεωρητικό και Τεχνικό Υπόβαθρο

Στο παρόν κεφάλαιο θα παρουσιαστούν και θα επεξηγηθούν ορισμένα βασικά θεωρητικά και τεχνικά στοιχεία από έννοιες που χρησιμοποιούνται για την υλοποίηση της παρούσης διπλωματικής εργασίας. Θεμέλιο και βάση όλων αποτελεί η τεχνική νοημοσύνη και η επιστήμη της μηχανικής μάθησης, της οποίας πιο συγκεντρωμένες εφαρμογές χρησιμοποιούνται σε βάθος, κυρίως και πρωτίστως η κατανόηση και επεξεργασια φυσικής γλώσσας (NLU & NLP - Natural Language Understanding/Processing). Παράλληλα θα συζητηθούν τυχόντες αλγόριθμοι, τεχνικές και μετρικές αξιολόγησης , όπως και εργαλεία προγραμματισμού - βιβλιοθήκες Python που χρησιμοποιήθηκαν κατά την εκπόνηση της διπλωματικής εργασίας.

3.1 Μηχανικη Μαθηση

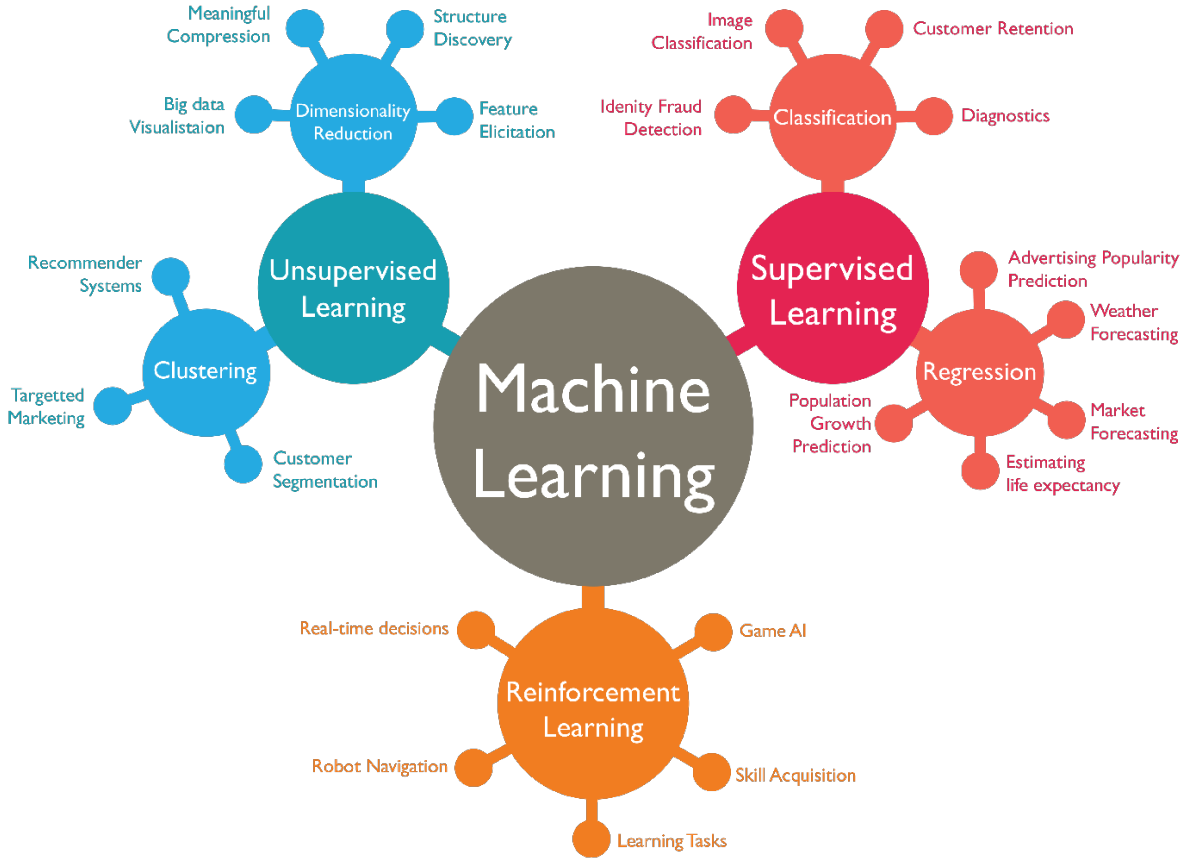
Η μηχανική μάθηση (Machine Learning - ML) αποτελεί σημαντικό τμήμα του τομέα της τεχνητής νοημοσύνης (Artificial Intelligence - AI), η οποία τα τελευταία χρόνια παρουσιάζει εντυπωσιακή εξέλιξη τόσο σε επίπεδο έρευνας, αλλά και σε αυτό της ευρύτερης υιοθεσίας και εφαρμογής στην αγορά. Βασικός στόχος της μηχανικής μάθησης είναι η βοήθεια στην επίλυση περίπλοκων και μεγάλων σε όγκο προβλημάτων. Αυτό επιτυγχάνεται με την εύρεση και αναγνώριση προτύπων - μοτίβων, βάσει μοντέλων που έχουν δημιουργηθεί - εκπαιδευτεί πάνω σε παρόμοιας φύσης δεδομένα. Αυτή η προσέγγιση με γνώμονα τα δεδομένα (Data-Driven Approach) επιτρέπει σε ένα σύστημα να αναγνωρίζει μια ιδιότητα σε κάποιο άγνωστο σε αυτό αντικείμενο ή να πραγματοποιεί έγκυρες σε μεγάλο βαθμό προβλέψεις χώρις πρότερη γνώση, μπορεί δηλαδή να "αποστάξει" την γενικότερη γνώση από τα δοθέντα δεδομένα και αμέσως να την εφαρμόσει για την επίλυση ενός παραπλησίου προβλήματος. Είναι συνεπώς μια μορφή τεχνητής νοημοσύνης η οποία αποσκοπεί στην εκκμάθηση από δεδομένα και να μην βασίζεται σε ρητούς προγραμματιστικούς κανόνες. Η ένταξη της μηχανικής μάθησης στην επιστήμη υπολογιστών ήταν

επαναστατική, καθώς επέτρεψε στους υπολογιστές να πραγματοποιούν υγιείς και αξιόπιστες προβλέψεις ή/και να επιλύουν τόσο βασικά όσο και περίπλοκα καθημερινά προβλήματα, τα οποία υπό κανονικές συνθήκες θα απαιτούσαν πολύτιμο χρόνο ή ανθρώπινο δυναμικό για να έλθουν εις πέρας. Μοντέλο μηχανικής μάθησης (Machine Learning Model) αποτελεί το αποτέλεσμα που προκύπτει από την εφαρμογή/εκπαίδευση ενός αλγόριθμου μηχανικής μάθησης, το σύνολο δηλαδή των μαθηματικών - στατιστικών εργαλείων που θα καθρισουν τον τρόπο με τον οποίο θα βρεθούν τα υπάρχοντα πρότυπα και μοτίβα στο σύνολο δεδομένων προορισμένο για εκπαίδευση, από το οποίο θα εκμαιεύτει η γενικευμένη γνώση του συστήματος. Το σύνολο αυτό δεδομένων ονομάζεται σετ εκπαίδευσης.

Κάποιοι εκ των βασικών τομέων στους οποίους δρηστηριοποιείται έντονα τόσο η έρευνα όσο και η εφαρμογή της μηχανικής μάθησης είναι οι εξής:

- Επεξεργασία Φυσικής Γλώσσας - Natural Languge Processing NLP
  - Ταξιονόμηση Κειμένου - Text Classification
  - Ανάκτηση Πληροφορίας - Information Retrival
  - Παραγωγή Κειμένου - Text Generation
  - Διαλογικα Συστήματα - Chatbots
  - Μετάφραση/Μεταγλώττιση Κειμένου - Text Translation
- Μηχανική Όραση - Computer Vision
  - Αναγνώριση Αντικειμένων/Προσώπων - Object/Facial Recognition
  - Εντοπισμός Αντικειμένων - Object Detection
- Αναγνώριση Ομιλίας - Speech Recognition
- Βελτιστοποίηση
  - Μηχανές Αναζήτησης - Search Engines
  - Εξατομίκευση - Personalization
  - Συστήματα Προτάσεων - Recommender Systems
- Ρομποτική
  - Αυτόνομα οχήματα - Autonomous/Self-driving Vehicles
- Εφαρμογές Επιστημών Υγείας - Healthcare & Medical Sciences Applications
- Κυβερνοασφάλεια - Cybersecurity
- Βιντεοπαιχνίδια - Video Games, Game AI

καθώς και πολλοί άλλοι τομείς ή/και συνδυασμοί αυτών, κάποιοι από τους οποίους φαίνονται στο σχήμα 3.1.



Σχήμα 3.1: Απεικόνιση τομέων και εφαρμογών Μηχανικής Μάθησης.
Πηγή: *https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications*

Βάσει των παραπάνω συμπεραίνεται ότι ένα κύριο προτέρημα της μηχανικής μάθησης είναι η δυνατότητα να εκπαιδεύεται με ελάχιστη ή καμία ανθρώπινη παρέμβαση, αλλά και ταυτόχρονα να βελτιώνονται. Οι αλγόριθμοι μηχανικής μάθησης κατά κύριο λόγο εντάσσονται σε τέσσερεις κατηγορίες:

- **Επιβλεπόμενη Μάθηση - Supervised Learning**: Η εκπαίδευση του μοντέλου μηχανικής μάθησης τελείται παρουσία ενός σετ εκπαίδευσης το οποίο περιέχει τόσο παραδείγματα της εισόδου αλλά και της επιθυμητης/σωστής εξόδου. Σε κάθε δηλαδή παράδειγμα/αντικείμενο στο σετ εκπαίδευσης έχει αποδοθεί μια ετικέτα - label που δρουν ως καθοδηγητήρια γραμμή για τον αλγόριθμο. Συνεπώς ο αλγόριθμος εντοπίζει σχέσεις μεταξύ των χαρακτηριστικών της εισόδου και την προκειμένη ετικέτα, δημιουργεί δηλαδή την σχέση αίτιου και αποτελέσματος στο σύνολο των μεταβλητών στην πληρότητα του σετ εκπαίδευσης. Με αυτόν τον τρόπο εντέλει εξάγεται ένας ευρύτερος γενικός κανόνας όπου σε μια παρόμοια είσοδο μπορεί να αποδοθεί το επιθυμητο αποτέλεσμα, μια απο τις ετικέτες. Συνήθως η επιβλεπόμενη μάθηση χρησιμοποιείται για την επιλύση προβλημάτων ταξινόμησης - classification ή παλινδρόμησης - regression (Αντιστοίχιση μιας εισόδου σε διακριτή ή συνεχή έξοδο αντίστοιχα).

- **Μη Επιβλεπόμενη Μάθηση - Unsupervised Learning**: Σε αντίθεση με την επιβλεπόμενη μάθηση, οι αλγόριθμοι μη επιβλεπόμενης μάθησης (Unsupervised

Learning) αναζητούν πρότυπα και μοτίβα σε σύνολα δεδομένων χωρίς προϋπάρχουσες ετικέτες (unlabelled data). Όπως επεξηγεί και το όνομα, η μη επιβλεπόμενη μάθηση χρήζει ελάχιστης ανθρώπινης επίβλεψης και επέμβασης. Σε μεγάλο βαθμό αλγόριθμοι μη επιβλεπόμενης μάθησης χρησιμοποιούνται για την επίλυση προβλημάτων ομαδοποίησης (clustering) και μείωσης διαστάσεων (dimensionality reduction).

- **Ημι-επιβλεπόμενη Μάθηση - Semi-supervised Learning**: Αλγόριθμοι ημιεπιβλεπόμενης μάθησης ή ασθενούς επίβλεψης (weak supervision) όπως υποδηλώνει και το όνομά τους, αποτελούν ένα μείγμα των 2 παραπάνω κατηγοριών, όπου δηλαδή, λαμβάνεται ως είσοδος ένα μικρό μέρος πληροφορίας το οποίο φέρει ετικέτες, αντίστοιχα με την επιβλεπόμενη μάθηση, αλλά σε συνδυασμό με μεγάλο όγκο δεδομένων άνευ ετικετών. Καθιστά προτιμότερη επιλογή για προβλήματα με τεράστιο όγκο δεδομένων καθώς παρακάμπτεται η δαπανηρή διαδικασία απόδοσης ετικετών στην πληρότητα τους, γεγονός ελαχιστοποιεί τον χρόνο και τους πόρους που χρειάζονται για την εκπαίδευση. Επίκαιρο παράδειγμα χρήσης ημι-επιβλεπόμενης μάθησης αποτελούν τα Μεγάλα Μοντέλα Γλώσσας - Large Language Models (LLMs). Αν γενικευθεί και απλουστευθεί αρκετά μπορεί να συμπεράνει κανείς ότι με τέτοιου είδους αλγόριθμους μάθησης μπορούν να επιλυθουν προβλήματα, ομαδοποίησης (clustering) και εν συνεχεία ταξινόμησης (clustering) ή ταξινόμισης ομάδων (cluster classification).

- **Ενισχυτική Μάθηση - Reinforcement Learning**: Η ενιχυτική μάθηση έχει βασικό στόχο την εκπαίδευση ενός αλγορίθμου πράκτορα - agent, ο οποίος εκπαιδεύεται με μέθοδο επιβράβευσης επιθυμητών και ευνοϊκών αποτελεσμάτων και "τιμωρίας" και αποθάρρυνσης των μη επιθυμητών αποτελεσμάτων αντίστοιχα. Σχετίζεται με και εμπνέεται, έως έναν βαθμό, από τον τρόπο που μαθαίνουν τα ανθρώπινα όντα, αλληλεπιδρώντας δηλαδή δυναμικά με το περιβάλλον.

Η μελέτη της παρούσας διπλωματικής εργασίας κατά κύριο λόγο πραγματεύεται έναν συνδυασμό προβλημάτων ταξινόμησης και ομαδοποίησης και συνεπώς χρησιμοποιούνται συνδυασμοί ημι-επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Δεν υπάρχει κάποιο πλέον αυστηρά ορισμένο σύνολο ετικετών πάνω στο οποίο θα πρέπει να προσαρμοστεί το σύνολο των δεδομένων.

3.2 Βαθια Μαθηση - Deep Learning

Η βαθιά μάθηση - Deep Learning (DL) αποτελεί ένα ακόμη πιο εξειδικευμένο υποσύνολο της μηχανικής μάθησης και κατ' επέκταση λοιπόν της τεχνητής νοημοσύνης. Η βαθιά μάθηση αποτελείται από αλγόριθμους που χρησιμοποιούν πολλαπλά στρώματα (layers) για να εντοπίσουν και να εξαγάγουν προοδευτικά πρότυπα, μοτίβα και χαρακτηριστικά από τα πρωτογενή δεδομένα. Αυτό το αποτέλεσμα επιτυγχάνεται με την χρήση Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) - Artifical Neural Networks (ANN/NN), τα οποία είναι ένας κλάδος μοντέλων μηχανικής μάθησης και

αποτελούνται από πολλούς κόμβους (nodes) που ονομάζονται τεχνητοί νευρώνες - artificial neurons και προσομοιώνουν τον τρόπο που λειτουργούν οι νεωρώνες στον ανθρώπινο εγκέφαλο. Κάθε σύνδεση νευρώνων όπως και οι συνάψεις στον εγκέφαλο έχουν την δυνατότητα να μεταφέρουν και να μεταβιβάζουν πληροφορία σε λοιπούς νευρώνες. Τα ΤΝΔ χωρίζονται σε 2 διακριτές κατηγορίες τοπολογίας: (α) Νευρωνικά Δίκτυα με Μετάδραση - Feeforward Neural Networks (FNNs) και (β) Νευρωνικά Δίκτυα με Ανατροφοδότηση - Recurrent Neural Networks (RNNs), όπου τα μεν (α) η πληροφορία μεταβιβάζεται ή ρέει προς μια μόνο κατεύθυνση χωρίς κυκλικές δομές ή βρόγχους και τα δε (β) έχουν την δυνατότητα ροής προς αμφότερες κατευθύνσεις με συνέπεια η έξοδος ενός νευρώνα να μπορεί να επηρεάσει την μελλοντική είσοδό του. Η βαθιά μάθηση έχει αποδειχθεί εξαιρετικά χρήσιμη και αποτελεσματική σε διεργασίες και προβλήματα αναγνώρισης εικόνας και κειμένων.

Η αρχιτεκτονική ενός βαθέος ΤΝΔ (βΤΝΔ) αποτελείται από το επίπεδο εισόδου, τα κρυφά επίπεδα και το επίπεδο εξόδου. Το επίπεδο εισόδου λαμβάνει τα δεδομένα και μεταβιβάζει την πληροφορία στο επόμενο επίπεδο, όπου αυτή επεξεργάζεται και επαναμεταβιβάζεται μέχρι το επίπεδο εξόδου, όπου παράγεται και εξάγεται η τελική απόφαση/πρόβλεψη του μοντέλου. Η εκπαίδευση ενός βΤΝΔ εμπεριέχει την διαδικασία της οπισθοδιάδοσης - backpropagation, κατά την οποία αναπροσαρμόζεται τα βάρη σε κάθε κρυφό επίπεδο αποσκοπώντας στην ελαχιστοποίηση του σφάλματος μεταξύ της αναμενόμενης και παραγώμενης εξόδου.

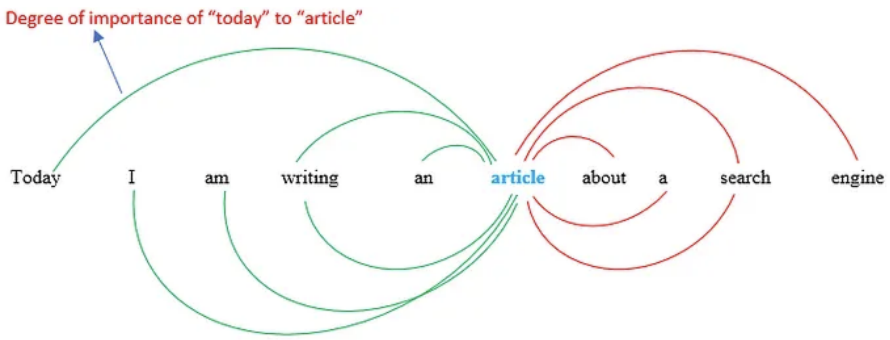
3.2.1 Ταξινόμηση Μηδενικής Βολής - Zero Shot Classification και Μεταφορική Μάθηση - Transfer Learning

Η ταξινόμηση μηδενικής βολής - zero shot classification αποτελεί ένα παράδειγμα εφαρμογή μεταφορικής μάθησης, δηλαδή η χρήση προεκπαιδευμένων μοντέλων για την επίλυση ενός προβλήματος διαφορετικής φύσεως από τα προβλήματα για τα οποία αρχικά προοριζόταν το μοντέλο. Συνήθως αυτό συμβαίνει λόγω της έλλειψης επαρκούς χαρακτηρισμένης με ετικέτα πληροφορίας - labelled data. Στην ταξινόμηση μηδενικής βολής το μοντέλο καλείται να ταξινομήσει ένα μικρό συνήθως κομμάτι κειμένου σε μια από τις δοθείσες ετικέτες χωρίς να έχει γίνει κάποιο επιπλέουσα εκπαίδευση ή προσαρμογή του μοντέλου. Για την συγκεκριμένη πρακτική είθισται να είναι πιο εύστοχα μεγάλα μοντέλα με υψηλό αριθμό (+100 εκατ.) παραμέτρων και στρωμάτων - layers. Αποτελεί ένα χρήσιμο εργαλείο για την ταξινόμηση κειμένου χωρίς την παράλληλη ύπαρξη ένος ευρύτερου συνόλουν κειμένων - corpus, όπως για παράδειγμα μια ερώτηση από έναν χρήστη σε μια μηχανή αναζήτησης ή σε ένα σύστημα ερωτοαπαντήσεων.

3.2.2 Μετασχηματιστές - Transformers

Οι μετασχηματιστές - transformers (TF) αποτελούν έναν εξειδικευμένο τύπο αρχιτεκτονικής και μεθόδου βαθιάς μάθησης για την επεξεργασία φυσικής γλώσσας, η οποία βασίζεται στην χρήση μηχανισμών προσοχής και πιο συγκεκριμένα αυτοπροσοχής - Self-Attention και προσοχή πολλαπλών κεφαλών - Multi-Head Attention.

Πρωτοεμφανίστηκαν το 2017 στην δημοσίευση ερευνητών της Google "Attention Is All You Need"[21] και έκτοτε αποτέλεσαν βασικό εργαλείο στην έρευνα και ευρύτερη εξέλιξη του τομέα. Η προοδευτική και καινοτόμα ενσωμάτωση των μηχανισμών αυτοπροσοχής δίνουν την δυνατότητα στα μοντέλα μετασχηματιστών να δίνουν βάρη και να σταθμίζουν την σημασία διαφορετικών στοιχείων - tokens ή λέξεων σε μία πρόταση και αυτά να συνυπολογίζονται στις προβλέψεις που γίνονται από το μοντέλο. Αυτό σε συνδυασμό με τη προσοχή πολλαπλών κεφαλών - multi head attention, που επιτρέπει στο μοντέλο να λαμβάνει υπόψιν πολλά μέρη της πρότασης εισόδου και συνεπώς να έχει καλύτερη κατανόηση του νοήματος και των συμφραζομένων, έχουν ως αποτέλεσμα οι εφαρμογές των μετασχηματιστών να παράγουν αξιοσημείωτα αποτελέσματα στην πλειοψηφία δοκιμών και εφαρμογών. Σε μεγάλο βαθμό αναπλήρωσαν τα Νευρωνικά Δίκτυα με Ανατροφοδότηση - Recurrent Neural Netwroks (RNNs) και τις εφαρμογές τους στα Δίκτυα Μακράς - Βραχείας Μνήμης - Long Short Term Memory Networks (LSTM)[22] τα οποία ήταν η βασική επιλογή για την επίλυση προβλημάτων που έχρηζαν κατανόησης φυσικής γλώσσας από τον υπολογιστή, αλλά αποδείχθηκαν οι αδυναμίες του σε σχέση με τους μετασχηματιστές στην δυσκολία τους και την ανάγκη πόρων για την επεξαργασία ακολουθιών μεγαλύτερων από λίγες προτάσεις. Η παρουσία πολλαπλών κεφαλών προσοχής σημαίνει ότι η κάθε κεφαλή μπορεί να εστιάζει σε διαφορετικό τμήμα και σημασιολογικό περιεχόμενο ενός κειμένου και συνεπώς να μπορούν να παράξουν η καθεμία μια ξεχωριστό αποτέλεσμα, τα οποία εντέλει θα συνδυαστούν για να σχηματιστεί η τελική αναπαράσταση του εκάστοτε εν λόγω κειμένου.



Σχήμα 3.2: Οπτικοποίηση Μηχανισμού Αυτοπροσοχής.

3.2.3 Αμφίδρομες Αναπαραστάσεις Κωδικποιητή από Μετασχηματιστές - Bidirectional Encoder Representations from Transformers (BERT)

Το 2019 παρουσιάζεται σε μία ερευνητική δημοσίευση από το εργαστήριο Google AI Language ένα μοντέλο για Αμφίδρομες Αναπαραστάσεις Κωδικοποιητή από Μετασχηματιστές - Bidirectional Encoder Representations from Transformers (BERT)[9], όπου αξιοποιείται οι αρχιτεκτονική από τους προαναφερθέντες μετασχηματιστές και χρησιμοποιούνται τεχνικές αμφίδρομης μάθησης για να επιτευχθεί πιο εις βάθως κατανόηση του γλωσσικού περιεχομένου. Τα μοντέλα BERT παράγουν διανυσματι-

κές αναπαραστάσεις κειμένου - text embeddings δίνοντας έμφαση και βαρύτητα σε σχέσεις γειτονικών λέξεων και των λοιπών συμφραζμένων, αντί απλά να αποδίδει αμιγώς ένα διάνυσμα σε κάθε μεμονωμένη λέξη του κειμένου. Χρησιμοποιούνται κωδικοποιητές μετασχηματιστών - Encoders from transformers σε οι οποίοι αποτελούνται από μια ακολουθία συμβόλων/λέξεων - token τα οποία στην συνέχεια αναπαραστούνται διανυσματικά και τίθενται προς επεξεργασία εντός νευρωνικού δικτύου. Η έξοδος του μοντέλου αποτελείται από μια ακολουθία διανυσματικών αναπαραστάσεων συμβόλων, με κάθε διάνυσμα να αντιστοιχεί σε ένα σύμβολο. Οι βασικές στρατηγικές εκπαίδευσης ενός BERT μοντέλου σε ένα σύνολο δισεκατομμυρίων λέξεων ήταν οι εξής δύο: (α) Αποκεκρυμμένη Γλωσσική Μοντελοποίηση - Masked Language Modelling (MLM), κατά τη οποία ένα ποσοστό των συμβόλων - token (της τάξεως του 15%) αντικαθίσταται από το σύμβολο [MASK] και βάσει των συμφραζομένων το μοντέλο προσπαθεί να επαναντικαταστήσει το σύμβολο [MASK] με το σωστό αρχικό σύμβολο και (β) Πρόβλεψη Επόμενης Πρότασης - Next Sentence Prediction (NSP), κατά την οποία στο μοντέλο δίνονται ως είδοσος δύο προτάσεις: A και B και αυτό καλείται να αξιολογήσει την υπόθεση ότι η πρόταση B έπεται της A. Η ευρύτερη φιλοσοφία και προσέγγυση των μοντέλων BERT έχει αποτελέσει βάσει για αρκετές παραλλαγές και βελτιώσεις μοντέλων όπως:

(α) Robustly Optimized BERT approach - RoBERTa[23], το οποίο παρήχθη από την ομάδα της Facebook AI, εκπαιδευμένο σε μεγαλύτερο σύνολο κειμένων - corpus και μεθόδους όπως δυναμική απόκρυψη - dynamic masking, περισσότερα σημεία ενδιαφέροντος και την έλλειψη της Πρόβλεψης Επόμενης Πρότασης (NSP), το σύνολο των οποίων το καθιστούν λιγότερο επιρρεπές στον θόρυβο και του επιτρέπουν να παράγει καλύτερα αποτελέσματα σε διάφορες τυποποιημένες δοκιμασίες - benchmarks σε σχέση με το BERT.

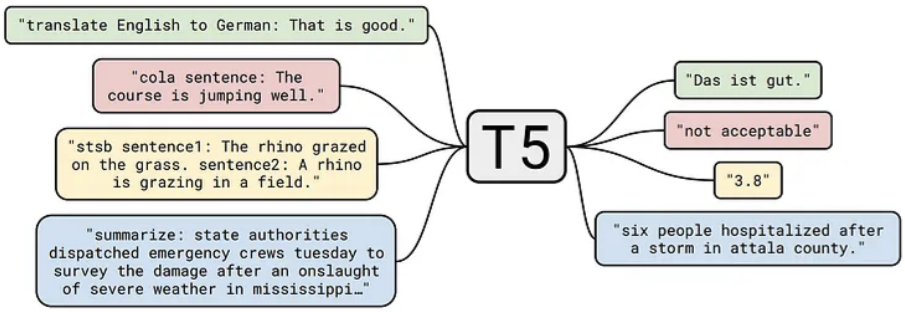
(β) A Lite BERT - ALBERT[24], το οποίο είναι ένα μεγαλύτερο μοντέλο του BERT το οποίο όμως χρησιμοποιώντας τεχνικές όπως διαστρωματικό μοίρασμα παραμέτρων - cross-layer parameter sharing και παραγοντοποίηση του διανυσματικού πίνακα - factorization of the embedding matrix μειώνεται σημαντικά ο αριθμός των απαιτούμενων από το μοντέλο παραμέτρων. Η τρίτη ειδοποιός διαφορά με το BERT είναι η χρήση της στρατηγικής πρόβλεψης σειράς προτάσεων - sentence-order prediction (OSP) αντί για NSP, κατά την οποία δίνεται έμφαση στην πρόβλεψης συνοχής μεταξύ δύο προτάσεων από το μοντέλο και όχι στην θεματική συσχέτιση.

(γ) DistilBERT[25], από ομάδα της HuggingFace, βασίζεται στην θεωρία της μεταφορικής μάθησης - transfer learning και συμπίεσης μοντέλου - model compression, όπου το μοντέλο μαθητής - student εκπαιδεύεται με σκοπό να αναπαράγει αποτελέσματα και συμπεριφορές ενός σαφώς μεγαλύτερου μοντέλου. Συγκεκριμένα το DistilBERT διατηρεί περίπου το 97% των γνωστικών ικανοτήτων του BERT μειώνοντας το μέγεθός του κατά 40%. Αντίστοιχης φιλοσοφίας μοντέλο είναι και το MiniLM[26], από ομάδα της Microsoft Research, του οποίου η αποσταγμένη γνώση βασίζεται σε διαφορετικά χαρακτηριστικά του διδάσκοντος μοντέλου - teacher model, και στις 2 περιπτώσεις το BERT BASE.

(δ) Bidirectional Auto-Regressive Transformer (BART)[27], αναπτύχθηκε και αυτό από ερευνητές της Facebook και στο οποίο η διαδικασία εκπαίδευσης αποτελείται από μία πιο περίπλοκη εκδοχή της αποκεκρυμμένης γλωσσικής μοντελοποίησης (MLM), όπου μια το κείμενο αλλοιώνεται από μια αυθαίρετη συνάρτηση θορύβου και στόχος είναι η εκπαίδευση ενός sequence-to-sequence μοντέλου που να δύναται να αποκαταστήσει το κείμενο.

3.2.4 Text-to-Text Transfer Transformer (T5) και Generizable T5-based Retrievers (GTR)[1]

Τα μοντέλα T5[28] βασίζονται και αυτά στην αρχιτεκτονική των μετασχηματιστών, με ταυτόχρονα κωδικοποιητές και αποκωδικοποιητές, συνολικά περιλαμβάνει 12 μπλοκ μετασχηματιστών και περιπου 220 εκατομμύρια παραμέτρους. Είναι εκπαιδευμένο στο C4 dataset (Colossal Clean Crawled Corpus)[29] το οποίο αποτελείται από 750GB αγγλικών κειμένων αντλημένα από το διαδίκτυο. Αντίστοιχα με το BERT χρησιμοποιεί και αυτό αποκεκρυμμένη γλωσσική μοντελοποίηση για την εκμάθησή του για την ορθή πρόβλεψη λέξεων. Χαρακτηριστική διαφορά δε αποτελεί το γεγονός ότι το T5 μπορεί να αποκρύπτει πολλαπλά σύμβολα - tokens ανά φορά σε αντίθεση με το BERT που αποκρύπτει σύμβολα ένα-ένα.



Σχήμα 3.3: Γραφικό παράδειγμα μοντέλου T5.

3.2.5 Διανυσματική Αναπαράσταση Κειμένου - Text Embedding

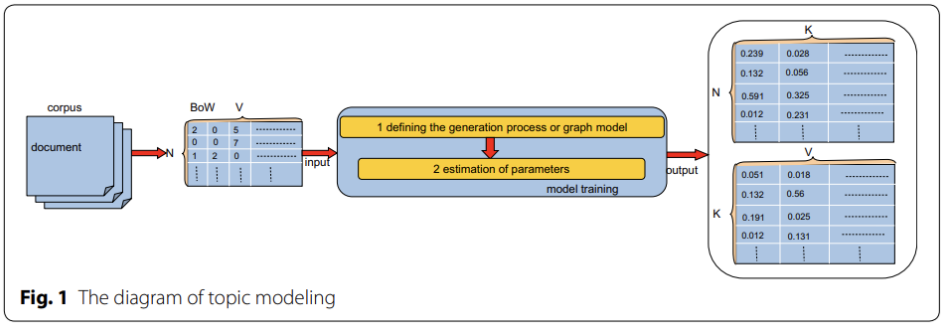
Το text embedding αποτελεί μια βασική τεχνική στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP), καθώς με αυτή τη διαδικασία ένα οποιοδήποτε πολυδιάστατο "αντικείμενο", στην προκειμένη περίπτωση κάποιο κείμενο, μπορεί να αναπαρασταθεί αριθμητικά από κάποια διανύσματα. Με αυτό τον τρόπο συνεπώς μπορούν τα παραγώμενα διανύσματα - text embeddings εν συνεχεία να επεξεργαστούν από αλγόριθμους μηχανικής μάθησης και συγκεκριμένα, γλωσσικά μοντέλα - language models. Βασικό χαρακτηριστικό και πρωτέρημα αυτής της τεχνικής είναι το γεγονός ότι τα παραγώμενα διανύσματα - text embeddings είναι σχεδιασμένα και υπολογισμένα ούτως ώστε να διατηρείται το σημασιολογικό περιεχόμενο και το νόημα των συμφραζομένων των λέξεων που αυτά αναπαριστούν. Χαρακτηριστικές

περιπτώσεις χρήσης τους, οι οποίες παρουσιάζονται στην παρούσα διπλωματική εργασία, είναι:

- Ταξινόμηση Κειμένου - Text Classification: όπου εκπαιδεύονται μοντέλα μηχανικής μάθησης και συνδέουν ετικέτες σε λέξεις, κείμενα και text embeddings. Στην συνέχεια στο κείμενο προς ταξινόμηση θα αποδοθεί μία ετικέτα, αύτη η οποία απέχει την χαμηλότερη απόσταση από το εκάστοτε κείμενο, μια διαδικασία σαφώς απλουστευμένη καθώς πλέον η μέτρηση απόστασης μπορεί να γίνει με απλά μαθηματικά εργαλεία, όπως η ευκλείδεια απόσταση.
- Σημασιολογική Αναζήτηση - Semantic Search: Μια ερώτηση ενός χρήστη τα κωδικοποιηθεί και θα αναπαρασταθεί από διανύσματα αντίστοιχα με το σύνολο κειμένων - corpus, μέσα στο οποίο βρίσκεται η απάντηση, συνεπώς όπως και παραπάνω, εάν υπολογιστεί η διανυσματική απόσταση του ερωτήματος του χρήστη και συγκριθεί με τα κείμενα, τα διανυσματικά πλησιέστερα είναι αυτα που είναι πιο πιθανό να περιέχουν την σωστή απάντηση.

3.2.6 Θεματική Μοντελοποίηση - Topic Modelling

Η θεματική μοντελοποίηση αποτελεί ένα στατιστικό εργαλείο μοντελοποίησης το οποίο χρησιμοποιείται για τον έλεγχο του βαθμού στον οποίο διάφορα θέματα εμφανίζονται σημασιολογικά σε έγγραφα εντός ενός συνόλου. Παράγονται βάρη για κάθε αρχείο του συνόλου, που αντικατοπτρίζουν πόσο προφανές είναι η καθεμία εκ των εντοπισμένων θεματολογιών σε αυτό. Η θεματική μοντελοποίηση αποτελεί μη επιβλεπόμενη προσέγγιση και βασίζεται στην "αλληλεπίδραση" και τις εσωτερικές σχέσεις των κειμένων του συνόλου. Δημιουργείται επίσης και ένα βαθμολογημένο σύνολο λέξεων, οι οποίες αποδίδουν στον μέγιστο βαθμό την εκάστοτε θεματολογία, τα πιο χαρακτηριστικά δηλαδή σύμβολα κάθε θέματος. Βασικές μέθοδοι και τεχνικές θεματικής μοντελοποίησης[19] είναι οι: (α) Λανθάνουσα Κατανομή Dirichlet - Latent Dirichlet Allocation (LDA), (β) Λανθάνουσα Σημασιολογική Ανάλυση - Latent Semantic Analysis (LSA), (γ) Μη-Αρνητική Παραγοντοποίηση Πινάκων - Non-Negative Matrix Factorization (NNMF) και (δ) BERTopic, που βασίζεται σε προσέγγιση βαθιάς μάθησης (DL approach).



Σχήμα 3.4: Διάγραμμα οπτικοποίησης ενός θεματικού μοντέλου.
Πηγή: Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus.[30]

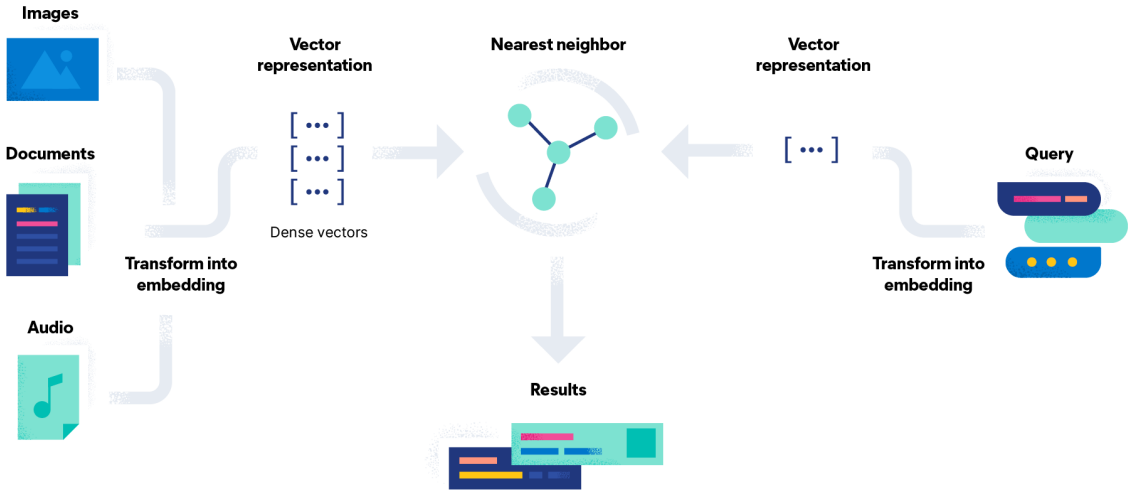
3.2.7 Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου - Term Frequency - Inverse Document Frequency (TF-IDF)

Η Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου (TF-IDF)[20] είναι μια μετρική που αποδίδει αριθμητικά την σημασία μιας λέξης σε ένα σύνολο κειμένων. Χρησιμοποιείται συχνά για την εξόρυξη πληροφοριών από κείμενα. Βασίζεται στον λόγο της συνότητας με την οποία εμφανίζεται ένας όρος σε ένα έγγραφο προς τον συνολικό αριθμό όρων σε στο έγγραφο (TF) και τον δεκαδικό ή φυσικό, ανάλογα με την εφαρμογή, λογάριθμο του λόγου του αριθμού των εγγράφων προς τον αριθμό των εγγράφων που περιέχουν τον όρο (IDF). Η μετρική TF-IDF υπολογίζεται από το γινόμενο των όρων TF και IDF. Βασικές περιπτώσεις στις οποίες χρησιμοποιείται η μετρική TF-IDF, όπως σε μηχανές αναζήτησης, περίληψη κειμένου, ταξινόμηση κειμένου και θεματική μοντελοποίηση. Μετρική αντίστοιχης φύσεως είναι και η Class Term Frequency - Inversed Document Frequency (c-TF-IDF), η οποία λαμβάνει υπόψη την ετικέτα / κλάση - class στην οποία ανήκει το έγγραφο. Στόχος είναι η αξιολόγηση της συνάφειας ενός όρου σε μία κλάση και η απόδοση της απαραίτητης σημασίας σε όρους μιας κλάσης εγγράφων, οι οποίοι βέλτιστα αποδίδουν σημασιολογικά το θέμα αυτής.

3.3 Συστηματα Ερωτοαπαντησεων - QA Systems

3.3.1 Σημασιολογική Αναζήτηση - Semantic Search

Για την λειτουργία ενός συστήματος ερωτοαπαντήσεων, η πιο σημαντική τεχνική ή διαδικασία είναι η σημασιολογική αναζήτηση - semantic search. Σε αντίθεση με μηχανισμούς αναζήτησης που χρησιμοποιούνται μέχρι και σήμερα, βάσει των οποίων αναζητούνται ακριβείς ταυτίσεις λέξεων ώστε να θεωρηθούν δύο ή περισσότερα κείμενα παρεμφερή, με την σημασιολογική αναζήτηση βασίζεται η σύγκριση στο νόημα και στο σημασιολογικό περιεχόμενο των κειμένων - semantics. Με την σημασιολογική αναζήτηση - Semantic Search θα χρησιμοποιηθούν τεχνικές μηχανικής και βαθιάς μάθησης για να μπορέσει ο υπολογιστής βέλτιστα να εκμαιεύσει το νόημα των κειμένων. Πιο συγκεκριμένα θα χρησιμοποιηθούν μοντέλα βαθιάς μάθησης για με στόχο να αναπαρασταθούν τα εν λόγω κείμενα διανυσματικά και να μπορέσουν στη συνέχεια να ταξινομηθούν και να καταταγούν με σειρά ομοιότητας ή συσχέτισης. Αυτή η τεχνική αναζήτησης γενικότερα ονομάζεται διανυσματική αναζήτηση - vector search. Βασικά πλεονεκτήματα της σημασιολογικής αναζήτησης υπέρ της αναζήτησης με λέξεις κλειδιά - keyword search, είναι ότι: (α) Μπορεί να εντοπίσει ομοιότητα και συσχέτιση μεταξύ κειμένων με κοινή θεματολογία αλλά πλήρως διαφορετική φρασεολογία και (β) μπορεί να αναγνωρίσει διαφορές σε αναζητήσεις με ίδιες μεν λέξεις, αλλά χρησιμοποιούμενες με διαφορετικό τρόπο, π.χ. “Milk chocolate” και “chocolate milk” δύο ίδιες λέξεις για μια αναζήτηση βάσει λέξεων κλειδιά, αλλά η διαφορές του συνδυασμού των οποίων μπορεί να γίνει αντιληπτός μόνο αν δοθεί βαρύτητα στην κατανόηση του σημασιολογικού περιεχομένου. Επίσης παρέχει την δυνατότητα, αν είναι εις βάθος γνωστά τα περιθώρια στα οποία και ο τρόπος με τον οποίο θα χρησιμοποιηθεί ένα σύστημα που αξιοποιεί σημασιολογική



Σχήμα 3.5: Περιγραφικό σχεδιάγραμμα για την διανυσματική αναζήτηση
Πηγή: *https://www.elastic.co/what-is/vector-search*

αναζήτηση, να παραμετροποιηθεί και να διαμορφωθεί ανάλογα με τις προδιαγραφές του χρήστη - personalization και να συνολογίζεται και αυτό κάθε φορά που γίνεται κάποια αναζήτηση.

3.3.2 Μοντέλα SQuAD (Stanford Question Answering Dataset)

Το σύνολο δεδομένων για ερωτοαπαντήσεις SQuAD[31] (Stanford Question Answering Dataset), είναι μια συλλογή ζευγών ερώτηση-απάντηση, κυρίως συλλεγμένα από άρθρα της Wikipedia. Η σωστή απάντηση δύναται να είναι κάποια ακολουθία από σύμβολα - tokens εμπεριεχόμενη στο εκάστοτε κείμενο. Το συγκεκριμένο σύνολο ερωτοαπαντήσεων, δημιουργήθηκε από μια ομάδα στο πανεπιστημιο Stanford, όπου πολλά άτομα με έμπιστη κρίση και γνωστική οξύτητα κλήθηκαν να απαντήσουν ερωτήσεις κατανόησης κειμένου και εφόσον δυνατό, να εντοπίσουν το τμήμα του κειμένου που περιέχει την απάντηση. Κατά αυτό τον τρόπο, το παρόν σύνολο δεδομένων - dataset έχει πολύ μεγάλο βαθμό διαφοροποίησης στις παραπάνω από 100,000 ερωτήσεις του. Ύστερα η ερευνητική ομάδα προσπάθησε να αποσαφηνίσει τις σχέσεις αιτίου-αιτιατού μεταξύ των ζευγών ερώτηση-απάντηση. Αφότου δημιουργήθηκε η πρώτη εκδοχή του SQuAD με αντίστοιχο μοντέλο προβλέψεων με μέτρια αποτελεσματα (51% F1 Score) σε αντίθεση με τις ανθρωπινες επιδόσεις: 86%, διάφορες ερευνητικές ομάδες παραμετροποίησαν ήδη υπάρχοντα μοντέλα βαθιάς μάθησης, ώστε να αποδίδουν καλύτερα σε προβλήματα αντίστοιχης φύσεως και σε benchmarks το μοντέλο χρησιμοποιείται στην εργασία αποδίδει περίπου αντίστοιχα με έναν άνθρωπο, βάσει του F1 score.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

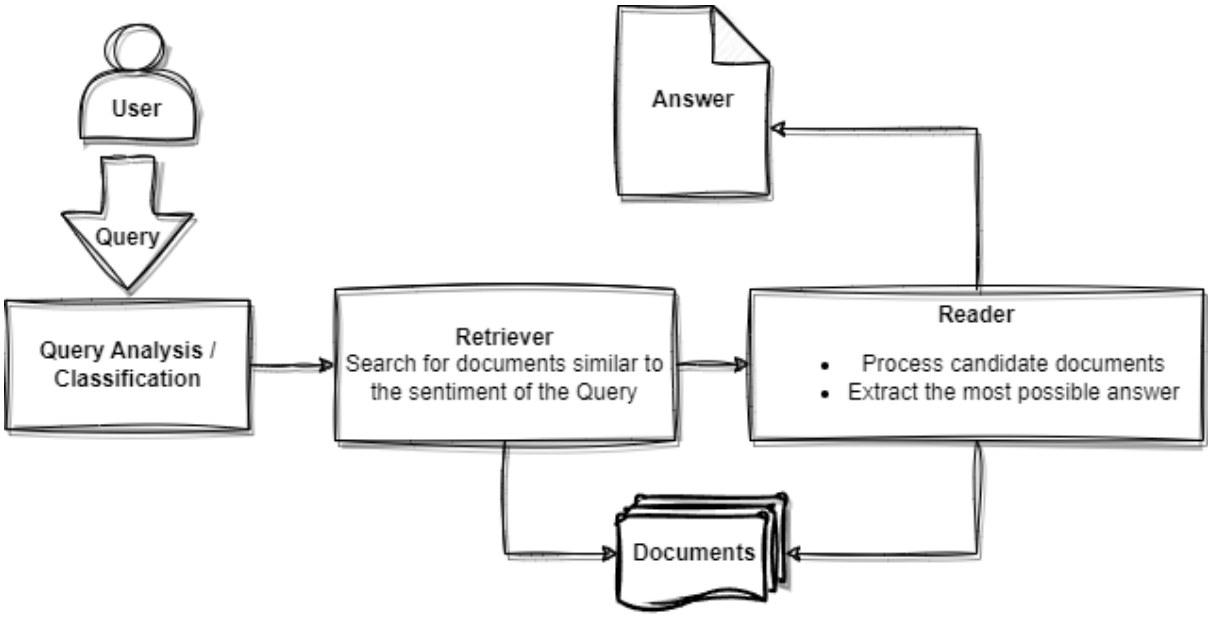
Σχήμα 3.6: Παράδειγμα SQuAD ζεύγους ερώτησης-απάντησης

3.4 Δομη Συστηματος Ερωτοαπαντησεων - QA

Κατα κύριο λόγο τα συστήματα ερωτοαπαντήσεων αποτελούνται από 3 βασικούς κόμβους - συστατικά μέρη: (α) το DocumentStore, το οποίο αποτελεί δομή τοσο για την οργανωμένη και δομημένη αποθήκευση εγγράφων, όσο και για την πηγή προσπέλασης αυτών από τα υπόλοιπα μέρη του συστήματος, (β) τον Retriever, ο οποίος είναι υπεύθυνος για να βρίσκει τα έγγραφα τα οποία είναι τα πλέον νοηματικά σχετκά με την εκάστοτε ερώτηση, επιτελεί συνεπώς την διαδικασία ανάκτησης πληροφορίας καθώς επιστρέφει τα πλέον πιθανά έγγραφα να περιέχουν την σωστή απάντηση και (γ) ο Reader ο οποίος είναι υπεύθυνος για την κατανόηση κειμένου, δηλαδή την διαδικασία κατά την οποία εντοπίζεται το συγκεκριμένο τμήμα του κειμένου, που περιέχει την απάντηση στην ερώτηση, έφοσον αυτό υπάρχει. [32]

3.4.1 Document Stores

Τα DocumentStores λειτουργούν ως βάσεις δεδομένων, τα οποία περιέχουν σε ορισμένη και δομημένη μορφή τα αρχεία κειμένου τα οποία έχουν ληφθεί από εξωτερική πηγή. Χάρις την ορισμένη μορφή την οποία έχουν τα αρχεία, αυτά γίνονται εύκολα προσπελάσιμα από τα υπόλοιπα μέρη του συστήματος.Συνεπώς τα έγγραφα πριν εισαχθούν σε ένα DocumentStore πρέπει να υποστούν προεπεξεργασία για να έρθουν στην κατάλληλη μορφή και να διαχωριστεί το περιεχόμενο τον κειμένων, από τυχούσα μεταπληροφορία - metadata, όπως τίτλο, συγγραφέα και διανυσματική αναπαράσταση - embedding, σε περίπτωση που αυτή χρειάζεται και αξιοποιείται από το είδος Retriver που χρησιμοποιείται. Υπάρχουν διάφορα είδη DocumentStore, κάποια εκ των οποίων χρησιμοποιούν βάσεις δεδομένων συγκεκρι-



Σχήμα 3.7: Σχηματική περιγραφή QA Συστήματος

μένου τύπου, όπως SQLite[1] και ElasticSearch[2]. Τα πιο δημοφιλή είδη DocumentStore είναι τα: InMemory, ElasticSearch, FAISS[3] και Milvus[4]

3.4.2 Ανάκτηση Πληροφορίας - Information Retrieval

Με τον όρο Ανάκτηση Πληροφορίας - Information Retrieval περιγράφετια η διαδικασία, κατά την οποία αναζητόνται έγγραφα σχετικά με την τεθείσα ερώτηση. Για αυτή την διαδικασία χρησιμοποιείται ο κόμβος Retriever από ένα σύστημα ερωτοαπαντήσεων. Υπάρχουν διάφορα είδη Retriever και χωρίζονται σε 2 ευρύτερες κατηγορίες, τους αραιούς - sparse και πυκνούς - dense, με ειδοποιό διαφορά μεταξύ τους το γεγονός ότι οι πυκνοί -dense αξιοποιούν βαθιά μάθηση για την ανάκτηση πληροφορίας - dense passage retrieval[33]. Αξιοποιούνται δηλαδή διανυσματικές αναπαραστάσεις τόσο των ερωτήσεων όσο και τον κειμένων με παρόμοια μοντέλα ώστε να υπολογισθεί ποια έγγραφα φέρουν την μέγιστη σημασιολογική ομοιότητα με την ερώτηση. Με αυτό τον τρόπο αποφεύγεται πιθανό σφάλμα που θα υπήρχε αν χρησιμοποιούνταν μέθοδοι bag-of-words όπως TF-IDF και εντάσσεται η σημασιολογική κατανόηση του κειμένου στην διαδικασία αυτή.

3.4.3 Κατανόηση Κειμένου - Reading Comprehension

Ο όρος Κατανόηση Κειμένου - Reading Comprehension αναφέρεται στην διαδικασία αναζήτησης της απάντησης σε μια ερώτηση μέσα σε ένα συγκεκριμένο κείμενο. Αυτη η διαδικασία γίνεται από τον κόμβο του Reader σε ένα σύστημα

[1] *https://www.sqlite.org/*
[2] *https://www.elastic.co/*
[3] *https://faiss.ai/*
[4] *https://milvus.io/*

ερωτοαπαντήσεων, στα κείμενα τα οποία θα επιστρέψει ο Retriever μετά την ανάκτηση πληροφορίας που θα εκτελέσει - Information Retrival. Για την διαδικασία αυτή εκτρελείται ανάλυση του κειμένου με ιδιαίτερη προσοχή σε λεπτομέρειες, σημασιολογικές αλλά και δομικές, όπως η σύνταξη του κειμένου, ώστε να βρεθεί το βέλτιστο τμήμα του κειμένου που παρέχει κατάλληλη απάντηση στην ερώτηση. Για αυτή την διαδικασία χρησιμοποιούνται ειδικώς παραμετροποιημένα μοντέλα Transformer, τα οποία έχουν εκπαιδευτει συγκεκριμένα στην διαδικασία κατανόησης κειμένου και απάντησης ερωτήσεων με σύνολα δεδομένων όπως το SQuAD[31], όπως φαίνεται και στο σχήμα 3.6

3.5 Εργαλεια Λογισμικου και Βιβλιοθηκες Python

3.5.1 Πακέτα και Βιβλιοθήκες Python

Pandas [5][34]: Βασικό εργαλείο για την αποθήκευση και ανάκληση δεδομένων για όλες τις διαδικασίες. Από την δημοφιλή βιβλιοθήκη pandas της Python κατά κύριο λόγο χρησιμοποιήθηκαν οι δομές **Pandas DataFrames**[6]. Τα DataFrames είναι δισδιάστατοι πίνακες δυναμικού μεγέθους που επιτρέπουν την αποθήκευση ετερογενών δεδομένων. Επιτρέπουν παράλληλα την έρευση των δεδομένων εντός αυτών βάσει δείκτη ευερτηρίου - index. Σε DataFrames αποθηκεύτηκαν όλα τα δεδομένα των συνόλων δεδομένων: όνομα και τίτλος εγγράφου, κείμενο, προεπεξεργασμένο κείμενο, ετικέτα - label, προβλεπόμενη ετικέτα - projected label και διανυσματική αναπαράσταση κειμένου embedding. Τέλος παρέχουν την ευκολία της υποστήριξης της αποθήκευσής τους σε αρχεία CSV (Comma Separated Values) και της ενσωματωμένης λειτουργίας οπτικοποίησης και απεικόνισης δεδομένων - data visualization.

Natural Language ToolKit (NLTK) [7][35] : Αποτελεί μια από τις βασικότερες βιβλιοθήκες για Επεξεργασία Φυσικής Γλώσσας στην Python. Στην παρούσα εργασία χρησιμοποιούνται σύνολα τελικών/τερματικών λέξεων - stop-words, σημείων στίξης και λημμάτων λέξεων αλλά και τα εργαλεία για την συμβολοποίηση και κωδικοποίηση του κειμένου - tokenizers.

PyYAML [8]: Αρχεία τύπου YAML (Yet Another Markdown Language) χρησιμοποιούνται για την εισαγωγή παραμέτρων σε κάθε στάδιο λειτουργίας από τον χρήστη, από τον καθορισμό του συνόλου δεδομένων και μοντέλων, μέχρι και την εισαγωγή των θεματολογιών του τελικού συστήματος.

transformers & sentence-transformers [9]: Είναι η επιμέρους βιβλιοθήκες της HuggingFace μέσω των οποίων εκτελούνται οι ξεχωριστές διαδικασίες των μοντέλων μετασχηματιστών, η διανυσματική αναπαράσταση των κειμένων και η ταξινόμηση μηδενικής

[5] *https://pandas.pydata.org/*
[6] *https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html*
[7] *https://www.nltk.org//*
[8] *https://pyyaml.org/wiki/PyYAMLDocumentation*
[9] *https://huggingface.co/docs/transformers/index*

βολής των κειμένων. Επίσης μέσω της βιβλιοθήκης sentence-transformers φορτώνονται στο σύστημα και τα μοντέλα που θα χρησιμοποιηθούν.

BERTopic [10]: Η βιβλιοθήκη που υλοποιεί, χρησιμοποιώντας μοντέλα και διανυσματικές αναπαραστάσεις από την "sentence-transformers", την θεματική μοντελοποίηση όλων των εγγράφων στο dataset. Εντός αυτής εκτελούνται τόσο οι αλγόριθμοι c-TF-IDF και ομαδοποίησης - Clustering όσο και μείωση διαστασιμότητας - dimensionality reduction. Διαθέτει επίσης απαραίτητες δυνατότητες οπτικοποίησης κάθε θεματικού μοντέλου.

Haystack [11]: Αποτελεί το θεμέλειο για τις λειτουργίες ερωτοαπαντήσεων - QA του συστήματος. Διαθέτει πολλά εργαλεία που επιτρέπουν την αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων - LLMs για την κατασκευή ξεχωριστών εφαρμογών/συστημάτων. Επιτρέπει την γρήγορη δοκιμή και χρήση εργαλείων NLP με ευκολία και ευελιξία. Στο συγκεκριμένο σύστημα, χρησιμοποιείται για την δημιουργία των συστημάτων QA και των ροών διεργασιών - pipelines καθώς και την ταξινόμηση και απάντηση ερωτησεων.

FastAPI [12]: Μια βιβλιοθήκη της Python που επιτρέπει την δημιουργία γρήγορων REST APIs με Python και χρησιμοποιείται για την επικοινωνία του συστήματος με κάθε επιμέρους σύστημα QA. Μέσω αυτής της βιβλιοθήκης δημιουργούνται API τερματικά - endpoints και POST μέθοδοι ώστε να ώστε μπορούν να ανταλλάσσονται δεδομένα μεταξύ ανεξάρτητων διεργασιών ή συστημάτων.

[10] *https://maartengr.github.io/BERTopic/index.html*
[11] *https://docs.haystack.deepset.ai/docs*
[12] *https://fastapi.tiangolo.com/*

4
Μεθοδολογία

Στο παρόν κεφάλαιο θα αναλυθεί η μεθοδολογία που ακολουθείται και υλοποιείται στην εργασία. Αντικείμενο σχολιασμού, περιγραφής και ανάπτυξης θα αποτελέσουν: (α) το σύνολο δεδομένων που χρησιμοποιήθηκε, (β) οι διεργασίες προεπεξεργασίας και καθαρισμού κειμένου, (γ) η θεματική μοντελοποίηση, (δ) η ταξινόμηση θεμάτων, (ε) η δημιουργία των συστημάτων ερωτοαπαντήσεων και (στ) η διαχείριση των ερωτήσεων του χρήστη, μαζί και με τεκμηριώσεις ως προς το γιατί χρησιμοποιείται ή προτιμάται το κάθε ένα από αυτά.

4.1 Γενικευμενη Ροη του Συστηματοσ

Όπως αναπτύχθηκε και προηγουμένως βασικός στόχος της παρούσας διπλωματικής είναι να δημιουργηθεί ένα σύστημα ερωτοαπαντήσεων - question-answering (QA) εξ' ολοκλήρου αυτόνομο και αποκεντρωμένο. Πρέπει να μπορεί δηλαδή το εν λόγω σύστημα να λάβει ως είσοδο έγγραφα κειμένου, να τα ταξινομήσει βάσει θεμάτων της προτιμήσεως του χρήστη - δημιουργού (developer) του συστήματος και εν συνεχεία να δημιουργηθούν τα ξεχωριστά υποσυστήματα ερωτοαπαντήσεων ορισμένου θέματος καθώς και ένας κεντρικός κόμβος - Master Node για την καθολική διαχείριση του τελικού συστήματος. Υπο αύτη την προσέγγιση θα δοκιμασθούν και θα αξιολογηθούν τυχόντα προνόμια υπολογιστικών ή αποθηκευτικών απαιτήσεων, ενώ ταυτόχρονα υπάρχουν εν γένει πλεονεκτήματα παραμετροποίησης και προσωπικοποίησης - personalization του συστήματος πάνω στις ανάγκες ή προτιμήσεις του χρήστη δημιουργού - developer. Παρακάτω θα αναλυθούν τα βήματα μεθοδολογίας που ακολουθήθηκαν για δημιουργηθεί και να δοκιμαστεί, αργότερα, ένα παρόμοιο σύστημα, τόσο από τα 2 στρώματα - layers ιεραρχικοποίησης στην ταξινόμηση και ομαδοποίηση εγγράφων έως την ακόλουθη δημιουργία και διαχείρηση μέσω τερματικών σημείων API των αποκεντρωμένων υποσυστημάτων ερωτοαπαντήσεων ορισμένου θέματος.

Ο χρήστης δημιουργός - developer επίλεγει τα θέματα στα οποία θέλει να κατανεμηθούν τα κείμενα και πάνω στα οποία να είναι βασισμένα τα αντίστοιχα επιμέρους υποσυστήματα QA και τα μοντέλα transformer τα οποία θα χρησιμοποιηθούν σε κάθε λειτουργία διανυσματικής απεικόνισης εγγράφων. Ύστερα δημιουργείται μέσω αλγορίθμων c-TF-IDF και της υποδομής BERTopic ένα θεματικό μοντέλο με κάθε κείμενο να περιγράφεται από ένα θέμα - topic μέσω των πιο σημαντικών λέξεων, μια λίστα λέξεων που επιλέγεται από τον αλγόριθμο c-TF-IDF. Ύστερα κάθε μια από αυτές τις λίστες που εκπροσωπούν ένα θέμα του μοντέλου ή ένα σύνολο κειμένων, ταξινομείται με την χρήση μιας ροής ταξινόμησης μηδενικής βολής - Zero-Shot Classification pipeline στα θέματα που έχει επιλέξει εξ' αρχής ο developer. Αναλόγως με το πού ταξινομηθεί το κάθε θέμα του μοντέλου, θα ταξινομηθεί και το κάθε έγγραφο που αυτό περιγράφει. Συνεπώς σε κάθε θεματολογία του χρήστη θα ταξινομηθούν έγγραφα με βάσει των πιο σημαντικών λέξεων και θα δημιουργηθεί ένα αρχείο που θα περιέχει όλη την έως τότε παρηγμένη πληροφορία ανά θεματολογία, κείμενο, προεπεξεργασμένο κείμενο και διανυσματική απεικόνιση.

Στην συνέχεια ο προαναφερθέντας κεντρικός κόμβος - Master Node, θα χρησιμοποιηθεί για να δημιουργήσει κάθε ένα από τα υποσυστήματα QA ορισμένου θέματος και ο ταξινομητής ερωτήσεων - query classifier, πάλι χρησιμοποιώντας αλγόριθμο ταξινόμησης μηδενικής βολής - zero-shot classfication. Αυτό συμπεριλαμβάνει την ροή αποδοχής και απάντησης ερωτήσεων - QA pipeline, αλλά και τις διόδους επικοινωνίας κάθε ενός από αυτά με τον κεντρικό κόμβο. Έτσι ο χρήστης καταναλωτής - consumer του συστήματος, όταν κάνει την ερώτησή του, αυτή θα ταξινομηθεί εντός τουταξινομητή ερωτήσεων και θα δρομολογηθεί μέσω των τερματικών σημείων API στο αντίστοιχο αρμόζον υποσύστημα για απάντηση. Όταν αυτή απαντηθεί θα αξιολογηθεί από τον Master Node η ευστοχία της ερώτησης και είτε θα επιστραφεί η απάντηση στον χρήστη ή θα επαναπροσπαθήσει το σύστημα να απαντήσει την ερώτηση πιο εύστοχα.

4.2 Σύνολα Δεδομένων - Datasets

Ως πρωτογενή δεδομένα και είσοδο στο σύστημα χρησιμοποιήθηκαν διάφορα ευρέως διαθέσιμα σύνολα δεδομένων - dataset κειμένων για Επεξεργασία Φυσικής Γλώσσας, όπως το πλέον δημοφιλές 20 Newsgroups, το AG News Dataset και το BBC News Archive Dataset, το οποίο εντέλει αποτέλεσε τον βασικό πυλώνα των δοκιμών. Επιλέχθηκαν τα παραπάνω σύνολα δεδομένων καθώς σε μεγάλο βαθμό περιείχαν αρθρογραφικά κείμενα ορισμένης θεματολογίας και μήκους μεγαλύτερου από 1-2 προτάσεις και παρείχαν παράλληλα ετικέτες - labelled data. Η παρουσία ετικετών και προταξινομημένων εγγράφων σημαίνει ότι με ευκολία μπορεί να ελεγχθεί εμπειρικά ή με την χρήση μετρικών η επίδοση και η ευστοχία/ακρίβεια του συστήματος. Κείμενα μεγάλου σχετικά μήκους, καθιστούν αρμόζουσα κάθε δοκιμή με την επιθυμητή και προβλεπόμενη χρήση ενός αντίστοιχου συστήματος, και προϊδεάζει για δυσκολίες υπολογισμού ή αστοχίες μοντέλων που έχουν αναπτυχθεί για σαφώς συντομότερα κείμενα. Παράλληλα σε κείμενα περισσότερων λέξεων μπορούν θεωρητικά να αναπτυχθούν πιο έντονα 1 ή περισσότερες θεματολογίες και αντικατοπτρίζουν ένα περιβάλλον θεματικής γνώσης αντίστοιχο με αυτό

μιας εγκυκλοπαίδειας. Αυτός ήταν ένας από τους λόγους που αποφεύχθηκαν σύνολα δεδομένων με δημοσιεύσεις του Twitter ή από κριτικές σε μέσα κοινωνικής δικτύωσης και διαδικτυακές υπηρεσίες, όπως κριτικές Airbnb κ.α.. Η πιο αυστηρά ορισμένη θεματολογία και έλλειψη θορύβου επίσης βοήθησε στον χαρακτήρα ερωτοαπαντήσεων του συστήματος, όπου είναι πιο εύκολο να ορισθούν δοκιμαστικές για το σύστημα ερωτήσεις. Πηγές των εν λόγω συνόλων είναι οι παρακάτω:

- 20 Newsgroups[13]
- AG News[14]
- BBC News Archive[15]

4.3 Προεπεξεργασια Κειμενου

Απαραίτητα στην επεξεργασία φυσικής γλώσσας είναι ορισμένα βήματα προεπεξεργασίας κειμένου ώστε να μπορέσουν τα μοντέλα μηχανικής μάθησης που θα εφαρμοστούν πάνω τους να αποστάξουν το μέγιστο του νοήματος και του σημασιολογικού περιεχομένου αυτών. Η διαδικασία αυτή, η οποία αλλιώς ονομάζεται καθαρισμός κειμένου - Text Cleaning, λειτουργεί ως στάδιο προετοιμασίας, όπου αφαιρούνται στοιχεία του κειμένου που δεν θα βοηθήσουν στον καθορισμό της θεματολογίας ή μπορεί να είναι "γκρίζα" σημεία αβεβαιότητας για ένα μοντέλο. Πιο συγκεκριμένα τα βήματα που ακολουθούνται είναι:

- **Συμβολοποίηση/Τμηματοποίηση - Tokenization του κειμένου**: Κατά αυτό το βήμα όλα τα στοιχεία ενός κειμένου λέξεις, φράσεις, προτάσεις, αριθμοί, σημεία στίξης κ.ο.κ. μετατρέπονται σε μεμονωμένα σύμβολα - tokens. Γίνεται δηλαδή πλήρης κατάτμηση του κειμένου σε όλα τα επιμέρους στοιχεία του. Αυτό επιτυγχάνεται με την χρήση ενός tokenizer, στην συγκεκριμένη περίπτωση: η συνάρτηση "word_tokenize()" από την βιβλιοθήκη Natural Language ToolKit - NLTK της Python.

- **Αφαίρεση των τελικών λέξεων - stop-words**: Τελικές λέξεις - Stop-words είναι λέξεις που εμφανίζονται συχνά στον γραπτό λόγο και συνήθως βοηθούν στην συνοχή, συνέχεια και ροή του λόγου χωρίς όμως να προσφέρουν ιδιαίτερη σημασιολογική αξία στο κείμενο. Άρθρα, αντωνυμίες, σύνδεσμοι, προθέσεις και προσδιορισμοί σε μεγάλο βαθμό θεωρούνται stop-words και αφαιρούνται από ένα κείμενο ώστε να διευκολύνουν την επεξεργασία και ταξινόμηση αυτού χωρίς να διαβρώνουν το νόημά του. Παραδείγματα τελικών λέξεων στην Ελληνική είναι: "ακόμη", "έχει", "στο", "εγώ", "οποιουδήποτε" και πολλές άλλες αντίστοιχες λέξεις. Στην προκειμένη περίπτωση χρησιμοποιείται το σύνολο stop-words πάλι από την βιβλιοθήκη Natural Language ToolKit - NLTK της Python.

[13] *http://qwone.com/~jason/20Newsgroups/*
[14] *https://huggingface.co/datasets/ag_news*
[15] *https://www.kaggle.com/datasets/hgultekin/bbcnewsarchive*

- **Αφαίρεση των σημείων στίξης**: Αντίστοιχα με τα stop-words αφαιρούμε από τα προς επεξεργασία κείμενα και τα σημεία στίξης ώστε να μην εισαχθούν στο μοντέλο ως ξεχωριστά σύμβολα, τα οποία σαφώς και δεν προάγουν το σημασιολογικό και θεματικό περιεχόμενο του κειμένου. Η αφαίρεσή τους γίνεται με το σύνολο "punctuation" της κλάσης "string" της Python με χρήση Regular Expressions για κάθε σύμβολο του κειμένου.

- **Λημματοποίηση - Lemmatization ή Στελεχοποίηση συμβόλων - Token Stemming**: Σημαντικό και πλέον ουσιαστικό στάδιο για την απλοποίηση του κειμένου αποτελεί ή λημματοποίηση, μια μέθοδος λεκτικής κανονικοποίησης - word normalization. Είναι μια διαδικασία κατά την οποία αναγάγεται κάθε λέξη στην πιο βασική μορφή της, το λήμμα/θέμα της λέξης. Αυτό συνεπάγεται την ταύτιση για τον υπολογιστή διαφόρων μορφών, κλίσεων ή χρόνων, ενός ρήματος ή και αντίστοιχα ουσιαστικών και επιθετικών προσδιορισμών. Δεδομένου ότι το βασικό νόημα και θέμα μιας πρότασης ή ενός κειμένου δεν έγκειται από τον χρόνο ή το πλήθος, αφαιρείται κατά αυτόν τον τρόπο σημαντικό κομμάτι πολυπλοκότητας, χωρίς όμως να μειώνουμε την ένταση της τυχούσης παρουσίας ενός θέματος. Προτιμάται ως διαδικασία καθαρισμούς από την στελεχοποίηση - stemming, καθώς αντιστοιχίζει κάθε λέξη που βρίσκεται στο λεξιλόγιο του Λημματοποιητη - Lemmatizer με το λήμμα της, σε αντίθεση με τους στελεχοποιητές - stemmers, οι οποίοι αποκόπτουν κάποιους από τους τελευταίους χαρακτήρες και ενδεχομένως να προκύπτουν λέξεις πλήρως άγνωστες, άντι έρθουν στην απλούστερη μορφή τους. Πάλι αντίστοιχα και με τον tokenizer και τα stop-words, χρησιμοποιείται ο λημματοποιητής από την βιβλιοθήκη NLTK της Python.

4.4 Θεματικη Μοντελοποιηση - Topic Modelling

Όπως αναπτύχθηκε παραπάνω με την θεματική μοντελοποίηση παράγονται διανυσματικές συσχετίσεις μεταξύ των εγγράφων στο σύνολο δεδομένων και παραγόμενων θεμάτων και αντίστροφα. Στόχος συνεπώς είναι ο αυτόματος εντοπισμός τόσο των προφανών όσο και λανθάνοντων υποκείμενων θεμάτων. Με αυτό τον τρόπο δίνεται η δυνατότητα εξωτερικευτούν και να εξαχθούν σχέσεις και ομαδοποιήσεις μεταξύ των κειμένων εντός ενός αχανούς συνόλου. Μπορεί να επιτευχθεί με στατιστική και πιθανολογική μοντελοποίηση - probabilistic modelling ή με μη επιβλεπόμενη μάθηση - unsupervised learning. Η ύστερη αποτελεί τη επιλογή για την παρούσα εργασία καθώς παρουσιάζει πολλές διαφορετικές επιλογές για κάθε προσέγγιση, αλλά και συγκεκριμένες ευκολίες για μετέπειτα βήματα της μεθοδολογίας του συστήματος. Οι πρώτες προσεγγίσεις θεματικής μοντελοποίησης, βασισμένες στην πιθανολογική μοντελοποίηση, όπως η Λανθάνουσα Κατανομή Dirichlet - LDA, Λανθάνουσα Σημασιολογική Ανάλυση - LSA και Μη-Αρνητική Παραγοντοποίηση Πινάκων - NNMF λαμβάνουν υπόψιν τους τα αρχεία ως σύνολα λέξεων χωρίς να δίνεται σημασία στην σειρά των λέξεων και επικεντρώνεται στην συχνότητα αυτών σε κάθε έγγραφο και κατ' επέκταση σε κάθε παραγόμενο θέμα. Συνεπώς είτε με την LDA όπου χρησιμοποιούνται απόλυτες συχνότητες λέξεων είτε με την NNMF,

όπου μπορεί να υλοποιηθεί η βαρύτητα κάθε όρου με TF-IDF, δεν υπάρχει επίγνωση των συμφραζομένων από το σύστημα. Αντιθέτως ο αλγόριθμος BERTopic καθώς χρησιμοποιεί διανυσματικές αναπαραστάσεις κειμένου - text embeddings μπορεί να εκμεταλλευτεί τα προτερήματα μοντέλων sentence-transformers και να παράξει την πιο εις βάθος γνώση.

Το BERTopic μετά την διανυσματική αναπαράσταση των κειμένων εφαρμόζει μείωση διάστασης όπως Uniform Manifold Approximation and Projection - UMAP ή Principal Component Analysis - PCA για να γίνει πιο εύκολη και λιγότερο απαιτητική η διαχείριση των κειμένων από το μοντέλο. Στην συνέχεια εφαρμόζονται αλγόριθμοι ομαδοποίησης, συμβολοποίησης και απόδοσης βαρύτητας στους όρους.

Στην προκειμένη περίπτωση το σύνολο διεργασιών στο BERTopic είναι το εξής:

- **Διανυσματική Αναπαράσταση - Text Embedding**:
  Για την παραγωγή διανυσματικών αναπαραστάσεων χρησιμοποιήθηκε, όπως αναφέρθηκε παραπάνω η υποδομή των sentence-transformers από την Hugging Face, καθώς παρέχει πλήρη συμβατότητα με την βιβλιοθήκη του BERTopic και παράλληλα προσφέρει ποιοτικά και εύστοχα αποτελέσματα. Πιο συγκεκριμένα χρησιμοποιήθηκαν ορισμένα συγκεκριμένα μοντέλα για την υλοποίηση: (α) "sentence-transformers/gtr-t5-base" και (β) "sentence-transformers/gtr-t5-large" τα οποία βασίζονται στα T5 μοντέλα και προτείνεται για την χρήση του σε πλαίσια σύγκρισης ομοιότητας προτάσεων - sentence similarity. "Χαρτογραφεί" προτάσεις ή παραγράφους σε διανυσματικό χώρο 768 διαστάσεων. (γ) "sentence-transformers/all-MiniLM-L6-v2" και (δ) "sentence-transformers/all-MiniLM-L12-v2" αντίστοιχα είναι δύο μοντέλα αναπτυγμένα σύμφωνα με την φιλοσοφία των MiniLM τα οποία όπως αναφέρθηκε στο θεωρητική αναδρομή βασίζονται στην απόσταγμένη γνώση απο BERT BASE μοντέλα. Είναι επίσης σχεδιασμένα για προβλήματα ομοιότητας προτάσεων και ομαδοποίησης - clustering και χρησιμοποιούναι διανυσματικό χώρο 384 διαστάσεων. Τα τέσσερα πρώτα μοντέλα χρησιμοποιούνται σε ζεύγη, μικρών-μεγάλων, ώστε παράλληλα με την φιλοσοφία του μοντέλου να αξιολογηθεί εμπράκτως και η διαφορά στις επιδόσεις των "μεγαλύτερων" εκδοχών τους με περισώτερα στρώματα - layers. (ε) "sentence-transformers/all-distilroberta-v1" παρόμοιας φιλοσοφίας με τα μοντέλα MiniLM, συνδυάζουν την απόσταξη γνώσης με την φιλοσοφία των RoBERTa μοντέλων, η οποία θεωρητικά αποφέρει καλύτερα αποτελέσματα σε διαδικασίες σημασιολογικού περιεχομένου. Ο διανυσματικός χώρος είναι 768 διαστάσεων και η χρήση του ενδείκνυται για αξιολόγηση ομοιότητας προτάσεων, ομαδοποίηση και σημασιολογική αναζήτηση - semantic search.

  Στην συγκεκριμένη περίπτωση θα γίνει χρήση κυρίως του "all-MiniLM-L6-v2" (γ) καθώς οι διανυσματικές του διαστάσεις ταιριάζουν με αυτές των μοντέλων που χρησιμοποιούνται αργότερα για τις λειτουργίες ερωτοαπαντήσεων και συνεπώς καθιστά τον επανυπολογισμό των διανυσματικών αναπαραστάσεων μη αναγκαίο, μειώνοντας έτσι τις συνολικές υπολογιστικές ανάγκες του

συστήματος. Ταυτόχρονα αποδίδει σε πολλές περιπτώσεις αντίστοιχα ή καλύτερα από τα υπόλοιπα μοντέλα και αυτό σε συνδυασμό με τα παραπάνω και την φιλοσοφία του, αυτή της αποσταγμένη μάθησης, γεγονός που του επιτρέπει να καταλαμβάνει λιγότερο αποθηκευτικό χώρο και να λειτουργεί γρηγορότερα αποτελούν λόγους για τους οποίους γίνεται η επιλογή αυτή.

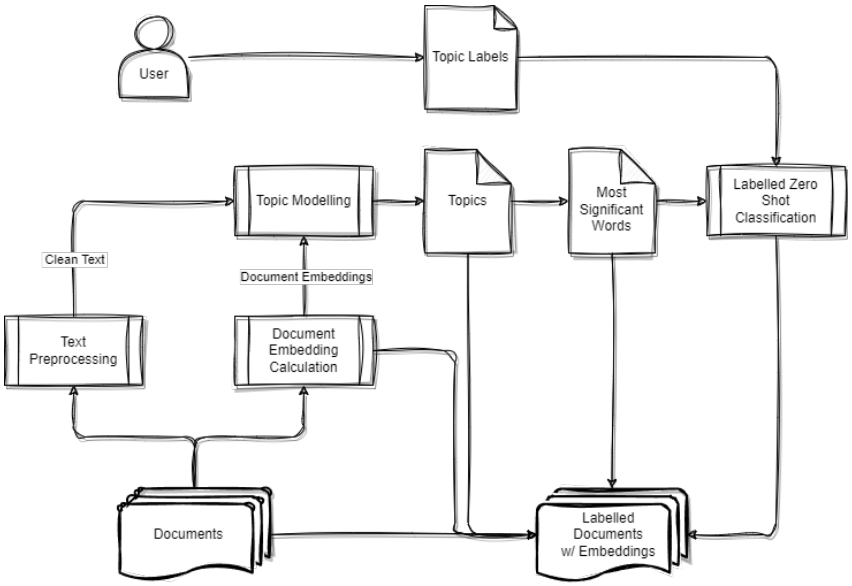
- **Μείωση Διαστασιμότητας - Dimensionality Reduction**:
  Η "Κατάρα της Διαστατιμότητας" συνεπάγεται την εν γένει δύσκολη διαχείρηση των υπάρχοντων δεδομένων και ειδικότερα σε μοντέλα ομαδοποίησης - cluster models. Συνεπώς σε μια προσπάθεια να "συμπυκνωθεί" η πληροφορία σε λιγότερες διαστάσεις χωρίς να αλλοιωθεί σημαντικά η πληροφορία, εφαρμόζονται τεχνικες όπως η Uniform Manifold Approximation and Projection (UMAP)[36] και η Principal Component Analysis (PCA)[37]. Στην προκειμένη περίπτωση χρησιμοποιείται η μέθοδος UMAP με στόχο να διατηρηθεί η γενικότερη δομή του συνόλου εγγράφων για να δημιουργηθούν εύστοχα ομάδες σημασιολογικά όμοιων εγγράφων.

- **Ομαδοποίηση Εγγράφων - Document Clustering**:
  Αφότου έχουν αναπαρασταθεί διανυσματικά τα κείμενα και έχει μειωθεί η διαστασιμότητα των δεδομέων χρησιμοποιείται μια εκ των δύο μεθόδων ομαδοποίησης: (α) k-Means[38], η οποία βασίζεται στην απόσταση από την μέση τιμή των ομάδων και (β) Hierarchical Density Based Spatial Clustering of Applications with Noise - HDBSCAN[39], η οποία βασίζεται στην πυκνότητα των παρατηρήσεων και δημιουργεί τις ανάλογες ομάδες. Με την HDBSCAN δύνανται να αναγνωριστούν και να δημιουργηθούν συστάδες ακανόνιστων σχημάτων και παράλληλα να εντοπισθούν ακραία σημεία - outliers.

- **Αναπαράσταση και Δημιουργία Θεμάτων - Topic Representation**:
  Πρωτού δημιουργηθούν τα θέματα, δημιουργούνται σύνολα λέξεων, ένα για κάθε ομάδα - cluster, τα οποία αποτελούνται από όλες τις λέξεις όλων των κειμένων στην ομάδα, εν τούτου παράγεται μια αναπαράσταση συνόλου-λέξεων - bag-of-words representation σε επίπεδο ομάδας και όχι σε επίπεδο εγγράφου ή dataset. Αυτό συμβαίνει ώστε να παραμένει ως έναν βαθμό ανεπηρέαστη από μεμονωμένα έγγραφα η διαδικασία δημιουργίας ενός διανύσματος των πιο αντιπροσωπευτικών λέξεων για κάθε μία ομάδα και να δοθεί έμφαση στις παραχθείσες ομάδες. Για την επίτευψη αυτού χρησιμοποιείται η προανφερθείσα παραλλαγή της TF-IDF, η c-TF-IDF, στην οποία κλάση θεωρείται κάθε cluster και βάσει του όρου c-TF-IDF υπολογίζονται οι πιο σημαντικές λέξεις του κάθε θέματος. Με αυτό τον τρόπο μειώνεται η σημασία μια λέξης που εμφανίζεται πολλές φορές σε πολλές κλάσεις καθώς συμπεραίνεται ότι δεν αποτελεί χαρακτηρηστικό ειδοποιούς διαφοράς μεταξύ θεμάτων.

4.5 Ταξινομηση Θεματων - Topic Classification

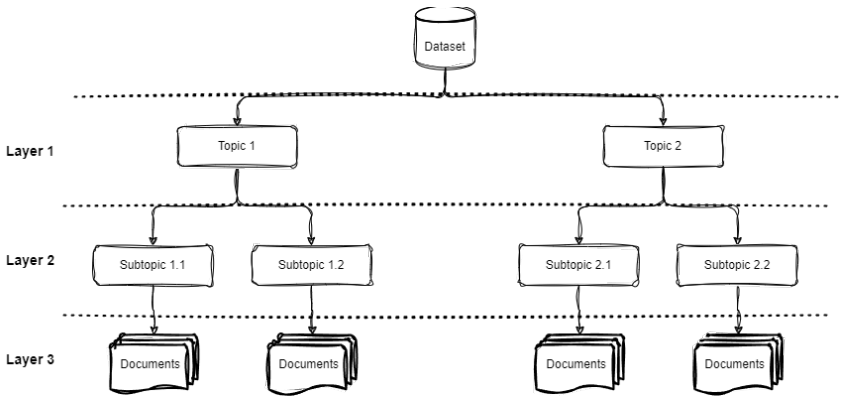
Θεωρείται δεδομένο ότι για την δημιουργία το συστήματος δίνεται ως είσοδος μια λίστα θεμάτων από τον χρήστη, το κάθε θέμα σε μορφή μίας λεξής που βέλτι-



Σχήμα 4.1: Σχέδιο Λειτουργίας του Υποσυστήματος Ιεραρχικής Ταξινόμησης Εγγράφων Βάσει Θέματος

στα περιγράφει το εν λόγω θέμα, π.χ. για ένα από τα θέματα τα οποία επιθυμούμε να περιέχουν δεδομένα σχετικά με την τεχνολογία και τα αθλητικά αντίστοιχα, θα είχε η λίστα θεμάτων - topic-label-list την εξής μορφή: [“technology”, “sports”] και φυσικά αντίστοιχα για λίστα περισσότερων θεμάτων. Μετά την θεματική μοντελοποίηση θα χρησιμοποιηθούν τα παρηγμένα διανύσματα των πιο σημαντικών λέξεων ως είσοδος σε μία ροή - pipeline της βιβλιοθήκης transformers της hugging face όπου εκτελείται ταξινόμηση μηδενικής βολής - zero shot classification με το μοντελο ”facebook/bart-large-mnli” με πιθανές ετικέτες labels τις λέξεις από το topic-label-list. Με αυτόν τον τρόπο, όπως περιγράφεται στο σχήμα 4.1 σε ένα αρχικά αταξινόμητο σύνολο εγγράφων - unlabelled dataset, μπορούν να αποδοθούν ετικέτες πιο φιλικές στην ανθρώπινη κατανόηση, σε μορφή κατηγορίας εγγράφου. Παράλληλα όμως δεδομένου του γεγονότος ότι πρώτα έχει υλοποιηθεί η θεματική μοντελοποίηση, καθιστάται δυνατό να μετατραπεί και να προσαρμοστεί σε αρκετά μεγάλο βαθμό το σύστημα, υπάρχει περιθώριο δηλαδή για εξατομίκευση - personalization και βελτιστοποίηση σε κάθε τυχούσα περίπτωση, βάσει τις απαιτήσεις του χρήστη και την φύση του προβλήματος. Χαρακτηριστικό παράδειγμα αποτελεί η δυνατότητα της διτής ή πολλαπλής ταξινόμησης, σε μία ομάδα εγγράφων (θέμα παρηγμένο από τον αλγόρθιμο του BERTopic) να αποδοθούν δηλαδή παραπάνω από μία ετικέτα. Αυτό συμβαίνει με τον ορισμό ενός κατώτατου κατωφλίου συσχέτισης, όπου αυτό μπορεί να είναι ένας δεκαδικός αριθμός που ανήκει (0,1) και θα εκφράζει την ελάχιστη αποδεκτή πιθανότητα που αποδίδει το μοντέλο σε μια ομάδα εγγράφων να ταιριάζει σε μια ετικέτα, ώστε να του αποδοθεί αυτή. Αυτή η διαδικασία ταξινόμησης είναι το τελευταίο στάδιο προετοιμασίας της οργάνωσης των εγγράφων πριν την δημιουργία συστημάτων ερωτοαπαντήσεων - QA systems, αλλά ταυτόχρονα όλα τα παραπάνω στάδια έχουν παράξει δεδομένα τα οποία καθιστούν το σύνολο δεδομένων - dataset σαφώς ταξινομημένο και πιο οργανωμένο.

Στο σχήμα 4.2 περιγράφεται σχηματικά η ιεραρχικότητα στην δομή του συστή-



Σχήμα 4.2: Περιγραφή Ιεραρχικότητας στην Οργάνωση των Αρχείων

ματος για την οργάνωση και ταξινόμηση των εγγράφων ανά θεματολογίες. Σαφώς το 3ο στρώμα - layer αποτελούν τα επιμέρους αρχεία αυτά καθαυτά, αλλά το 1ο στρώμα - layer είναι οι ευρύτερες θεματολογίες ορισμένες από τον χρήστη σε μορφή μονολεκτικών ετικετών, την λίστα θεμάτων - topic-label-list και τέλος το 2ο στρώμα - layer αποτελείται από τα παραγόμενα θέματα από την θεματική μοντελοποίηση με BERTopic. Συνεπώς δημιουργούνται ομάδες - clusters κειμένων με κοινή θεματολογία που περιγράφονται από ένα σύνολο λέξεων, οι πιο σημαντικές λέξεις ανά θέμα, με c-TF-IDF, και η τελική ταξινόμηση των εγγράφων προκύπτει από την ταξινόμηση ή αντιστοίχιση των subtopics του layer 2 με τα καταλληλότερα topics από το layer 1. Με αυτό τον τρόπο λοιπόν πέρα από την ενδελεχή οργάνωση του συνόλου κειμένων επιτυγχάνεται και μια ιεραρχική δομή στα δεδομένα.

4.6 Δημιουργια Συστηματων Ερωτοαπαντησεων

Επόμενο στάδιο αποτελεί η δημιουργια του ευρύτερου συστήματος ερωτοαπαντήσεων, όπως εμφανίζεται στο σχήμα 4.3, του οποίου οι αρμοδιότητες θα είναι οι εξής 3: (α) Να δημιουργήσει το κάθε υποσύστημα ερωτοαπαντήσεων ορισμένου θέματος βάσει ενός αρχείου διαμόργωσης YAML και να αρχικοποιηθεί η σύνδεση τους με τον βασικό κόμβο - master node μέσω API Endpoints, (β) Να δέχεται ερωτήσεις από τον χρήστη και να τις ταξινομεί σε κάποιο από τα δοσμένα labels του topic-label-list ορισμένο από τον χρήστη και (γ) να βρίσκει και να επιστρέφει την ή τις πιο σωστές και ταιριαστές απαντήσεις από το σύνολο των εγγράφων. Για την υλοποίηση του τμήματος αυτού θα χρησιμοποιηθεί η υποδομή και τα εργαλεία της βιβλιοθήκης farm-haystack από την deepset για την Python. Μέσω της Haystack θα χρησιμοποιηθούν όλα τα απαραίτητα εργαλεία για την είσοδο και προετοιμασία των εγγράφων για ερωτοαπάντηση, την ταξινόμηση ερωτήσεων και ανάκτηση και επιστροφή απαντήσεων. Σαφής και ειδοποιός διαφορά μεταξύ ενός "απλού", run-of-the-mill, QA συστήματος και του προτεινόμενου στην παρούσα εργασία, είναι το γεγονός ότι έχει προηγηθεί η προεπεξεργασία και θεματική μοντελοποίηση των εγράφων και συνεπώς στόχος είναι τα παραγώμενα συστήματα να είναι αυτομάτως συγκεκριμένης θεματολογίας - domain specific, να κατέχουν και να ανακτούν δηλαδή γνώση σχετική μόνο με το ευρύτερο θέμα πάνω στο οποίο είναι βασισμένα.

4.6.1 Εισαγωγή στα Συστήματα Ερωτοαπαντήσεων και στην βιβλιοθήκη Haystack

Η δομή και λειτουργία των παραπάνω συστημάτων βασίζονται στην σημασιολογική αναζήτηση - semantic search, η οποία διαφέρει με διαφορετικές τεχνικές αναζήτησης κειμένου, στις οποίες αναζητείται η ταύτιση λέξεων μεταξύ ερώτησης και απότελεσμάτων. Αντίθετα η σημασιολογική αναζήτηση βασίζεται στην διανυσματική αναζήτηση - vector search για να αξιολογήσει τις πιθανές απαντήσεις σε μια ερώτηση. Κωδικοποιούνται και τα δύο άκρα του ζεύγους ερώτηση-απάντηση και μετατρέπονται σε σημασιολογικά αναζητήσιμες οντότητες και εν συνεχεία συγκρίνονται τα διανύσματα, ώστε να κατανεμηθούν από πιο πιθανή σε λιγότερο πιθανή απάντηση στην εκάστοτε ερώτηση.

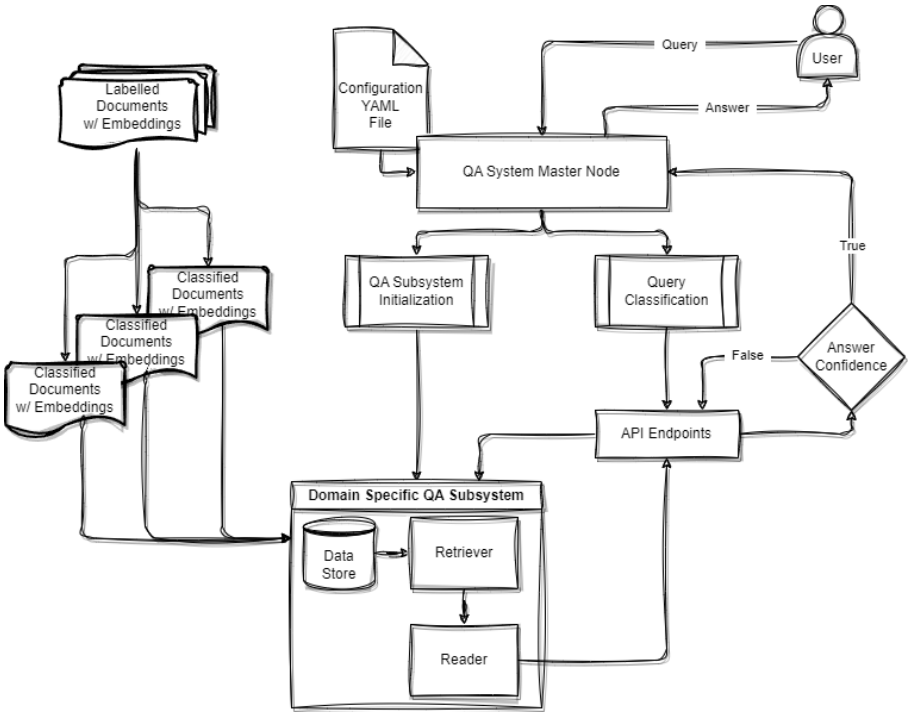
Κάθε επιμέρους σύστημα για να λειτουργεί αποτελεσματικά χρειάζεται τουλάχιστον τους τρεις παρακάτω κόμβους: (α) ένα Document Store, όπου θα αποθηκευτούν με δομημένο τρόπο όλα τα αρχεία, (β) έναν Retriever, ο οποίος θα χρησιμοποιηθεί για την σύγκριση της ερώτησης με τα αρχεία του Document Store και θα επιστρέψει τα κατανεμημένα αρχεία, (γ) ένας Reader, ο οποίος θα δεχτεί ως είσοδο τα εξαγώμενα αρχεία από τον Retriever και θα εντοπίσει στα εν λόγω αρχεία τα σημεία στα οποία βρίσκονται οι απαντήσεις. Πιο συγκεκριμένα:

- **Document Store**:
  Ένα Document Store είναι ένας αποθηκευτικός μηχανισμός που χρησιμοποιείται από της υποδομή της Haystack για την αποθήκευση και διαχείριση εγγράφων. Τα έγγραφα σε ένα Document Store αποθηκεύονται με μορφή ενός λεξικού - dictionary της Python, το οποίο περιέχει τόσο το κείμενο του εγγράφου όσο και ό, τι επιπλέον στοιχείο είναι διαθέσιμο για το έγγραφο, metadata, όπως ετικέτα από την ταξινόμηση προηγουμένως. Τα Document Stores είναι υψηλής σημασίας για την διαδικασίας της σημασιολογικής αναζήτησης και κατ' επέκταση την απάντησης ερωτήσεων, καθώς αποτελούν το μέσο για γρήγορη και αποδοτική προσπέλαση και ανάκτηση των πληροφοριών από τα έγγραφα. Στην υποδομή της Haystack είναι διαθέσιμοι διάφοροι τύποι Document Stores, όπως elasticSearch, FAISS, SQL και InMemory Document Store. Στην προκειμένη περίπτωση χρησιμοποιούνται Document Stores τοπικής μνήμης, InMemoryDocumentStore και στο λεξικό, dictionary του καθενός, θα είναι το περιεχόμενο του κειμένου, το indexing από την Haystack και η ετικέτα του από το Zero Shot Classification του εγγράφου. Σκοπός είναι να δημιουργείται ένα ξεχωριστό Document Store για κάθε ένα από τα topic labels, ένα για κάθε επιμέρους σύστημα ερωτοαπαντήσεων. Με αυτό τον τρόπο μπορεί να χρησιμοποιηθεί οποιοσδήποτε τύπος Document Store κρίνεται βέλτιστος για το εκάστοτε σύστημα, λαμβάνοντας υπόψη την φύση και περιβάλλον λειτουργίας του και την συνέργειά του με το υπόλοιπο σύστημα και παράλληλα να ελαχιστοποιεί θεωρητικά τον φόρτο των Retriever και Reader, καθώς θα πρέπει να ανατρέξουν σε σαφώς λιγότερα αρχεία στην αναζήτησή τους για πιθανές απαντήσεις.

- **Retriever**:
  Ο κόμβος του Retriever είναι υπεύθυνος για την αναδρομή σε όλα τα αρχεία

που περιέχονται στο Document Store και την σύγκριση των διανυσματικών αναπαραστάσεων των εγγράφων με αυτή της κάθε ερώτησης. Έξοδος του Retriever είναι ένα set αρχείων που είναι σημασιολογικά παραμφερή με την ερώτηση. Όπως και με τα Document Stores υπάρχουν διάφοροι τύποι από Retriver. Ο πλέον συνηθισμένος θεωρείται ο BM25 Retriever[40], ο οποίος φέρνει εις πέρας την σύγκριση των αρχείων χωρίς να χρησιμοποιεί νευρωνικά δίκτυα, αντ' αυτου χρησιμοποιεί την συνάρτηση ταξινόμησης Okapi BM25, μια παραλλαγή του TF-IDF. Στην παρούσα εργασία, όμως λόγω του γεγονότος ότι χρησιμοποιούμε embeddings από sentence-transformers μοντέλα μπορούν να αξιοποιηθούν τα ήδη υπάρχοντα embeddings, χωρίς να απαιτείτια επιπλέοντες υπολογιστικοί πόροι. Αυτό είναι δυνατό με την χρήση EmbeddingRetriever, όπου η διαδικασία της σύγκρισης του σημασιολογικού περιεχομένου γίνεται με την χρήση των embeddings από το μοντέλο της επιλογής μας, δηλαδή αυτό με το οποίο έγινε η θεματική μοντελοποίηση στα προηγούμενα βήματα Και το οποίο θα καθοριστεί βάσει δοκιμών στο επόμενο στάδιο, άλλα δίνεται μεγαλύτερη έμφαση σε μοντέλα με διανυσματικές διαστάσεις n=384, καθώς ταιριάζουν χωρικά και με τα μοντέλα των Readers. Συνεπώς όταν καταχωρηθούν τα αρχεία στο Document Store στα εντός των metadata για κάθε αρχείο θα είναι και το embedding του. Ο αριθμός των εγγράφων που θα επιστρέψει η συνάρτηση retrieve του Retriever, εξαρτάται από την μεταβλητή top_k, οριζόμενη από τον χρήστη και ίση με 10 από προεπιλογή. Η βέλτιστη τιμή βέβαια για την μεταβλητή top_k εξαρτάται από τον αριθμό και την διαφοροποίηση των αρχείων στο Document Store, αν είναι πολλά και παρεμφερή αρχεία, τοτε επιλέγεται μεγαλύτερος αριθμός εγγράφων.

- **Reader**:
  Η θέση του κόμβου Reader έπαιται αυτή του Retriever και δέχεται ως είσοδο τα αρχεία που επιστρέφει ο Retriever. Εν συνεχεία στα αρχεία αυτά θα αναζητηθεί από τον Reader το τμήμα του κειμένου που πιο εύστοχα απαντάει στην ερώτηση που έχει γίνει. Η αναζήτηση αυτή γίνεται πάλι με την χρήση μοντέλου μετασχηματιστών - transformers, αλλά για την προκειμένη χρήση είναι απαραίτητο να χρησιμοποιηθεί ένα μοντέλο τροποποιημένο - fine-tuned με σκοπό τις ερωτοαπαντήσεις - question answering (QA). Η τροποποίηση εν προκειμένου είναι το Stanford Question Answering Dataset - SQuAD fine-tuning. Από τους διαθέσιμους τύπους Reader από την Haystack, η επιλογή μας είναι οι FARM Readers καθώς είναι οι βέλτιστα σχεδιασμένοι για να λειτουργούν με μοντέλα βαθιάς μάθησης και ταυτόχρονα με την υποδομή της Haystack. Το μοντέλο που θα χρησιμοποιηθεί είναι το "deepset/reoberta-base-squad2" καθώς δεν είναι εφικτό να χρησιμοποιηθούν τα embeddings που έχουν υπολογιστεί στα προηγούμενα βήματα. Σημειοταίο είναι το γεγονός ότι το έργο του Reader είναι το σαφώς πιο απαιτητικό από άποψη υπολογιστικής δυσκολίας και ανάγκης υπολογιστικών πόρων και συνεπώς, γίνεται βέβαιο ότι μπορεί να αξιοποιηθεί κάποιου είδους μονάδα επεξεργασίας γραφικών - GPU για την επιτάχυνση των υπολογισμών, κάτι που υποστηρίζεται από την Haystack.



Σχήμα 4.3: Σχέδιο Λειτουργίας του Συστήματος Διαχείρισης Υποσυστημάτων και Ερωτήσεων

Επόμενο στάδιο αποτελεί η σύνδεση των παραπάνω κόμβων μεταξύ τους σε μία μια ροή διαδικασιών - ενεργειών, ένα pipeline. Για την συγκεκριμένη περίπτωση συμπίπτει ο τρόπος λειτουργίας του ExtractiveQAPipeline, το οποίο έχοντας πρόσβαση σε ένα Document Store συνδυάζει την λειτουργία ενός Retriever και ενός Reader για να βρει και να επιστρέψει την απάντηση σε μια ερώτηση, καθώς και το σημείο πυ βρίσκεται αυτή στο κείμενο. Εναλλακτική αποτελεί επίσης το GenerativeQAPipeline όπου συνδυάζεται ένας Retriver με έναν Generator ώστε με βάση την πληροφορία από τα αρχεία που επιστρέψει ο Retriever η έξοδος του Pipeline να είναι εξ' ολοκλήρου παραγώμενη από ένα παραγωγικό μοντέλο τεχνητής νοημοσύνης - generative AI model, διαδικασία όμωςπου απαιτεί περισσότερους υπολογιστικούς πόρους και συνδρομές σε εξωτερικές υπηρεσίες.

```
for i, topic in enumerate(topic_labels):
  temp = {
    topic:{
        'dataframe_filepath': f'./lib/classified/{topic}_document',
        'embedding_model': embedding_model,
        'reader_model': reader_model,
        'retriever_top_k': retriever_top_k,
        'reader_top_k': reader_top_k,
        'port_number': int(9001 + i)
        }
    }
```

Σχήμα 4.4: Βρόγχος "for" για την δημιουργία επιμέρους YAML config αρχείων

Για την δημιουργία των επιμέρους συστημάτων, θα αποθηκευτεί ένα pandas DataFrame για κάθε θέμα, το οποίο θα περιέχει όλη την σχετική πληροφορία για

```yaml
business:
  dataframe_filepath: ./lib/classified/business_document
  embedding_model: sentence-transformers/all-MiniLM-L6-v2
  port_number: 9001
  reader_model: deepset/tinyroberta-squad2
  reader_top_k: 5
  retriever_top_k: 10

entertainment:
  dataframe_filepath: ./lib/classified/entertainment_document
  embedding_model: sentence-transformers/all-MiniLM-L6-v2
  port_number: 9002
  reader_model: deepset/tinyroberta-squad2
  reader_top_k: 5
  retriever_top_k: 10

politics:
  dataframe_filepath: ./lib/classified/politics_document
  embedding_model: sentence-transformers/all-MiniLM-L6-v2
  port_number: 9003
  reader_model: deepset/tinyroberta-squad2
  reader_top_k: 5
  retriever_top_k: 10

sport:
  dataframe_filepath: ./lib/classified/sport_document
  embedding_model: sentence-transformers/all-MiniLM-L6-v2
  port_number: 9004
  reader_model: deepset/tinyroberta-squad2
  reader_top_k: 5
  retriever_top_k: 10

tech:
  dataframe_filepath: ./lib/classified/tech_document
  embedding_model: sentence-transformers/all-MiniLM-L6-v2
  port_number: 9005
  reader_model: deepset/tinyroberta-squad2
  reader_top_k: 5
  retriever_top_k: 10
```

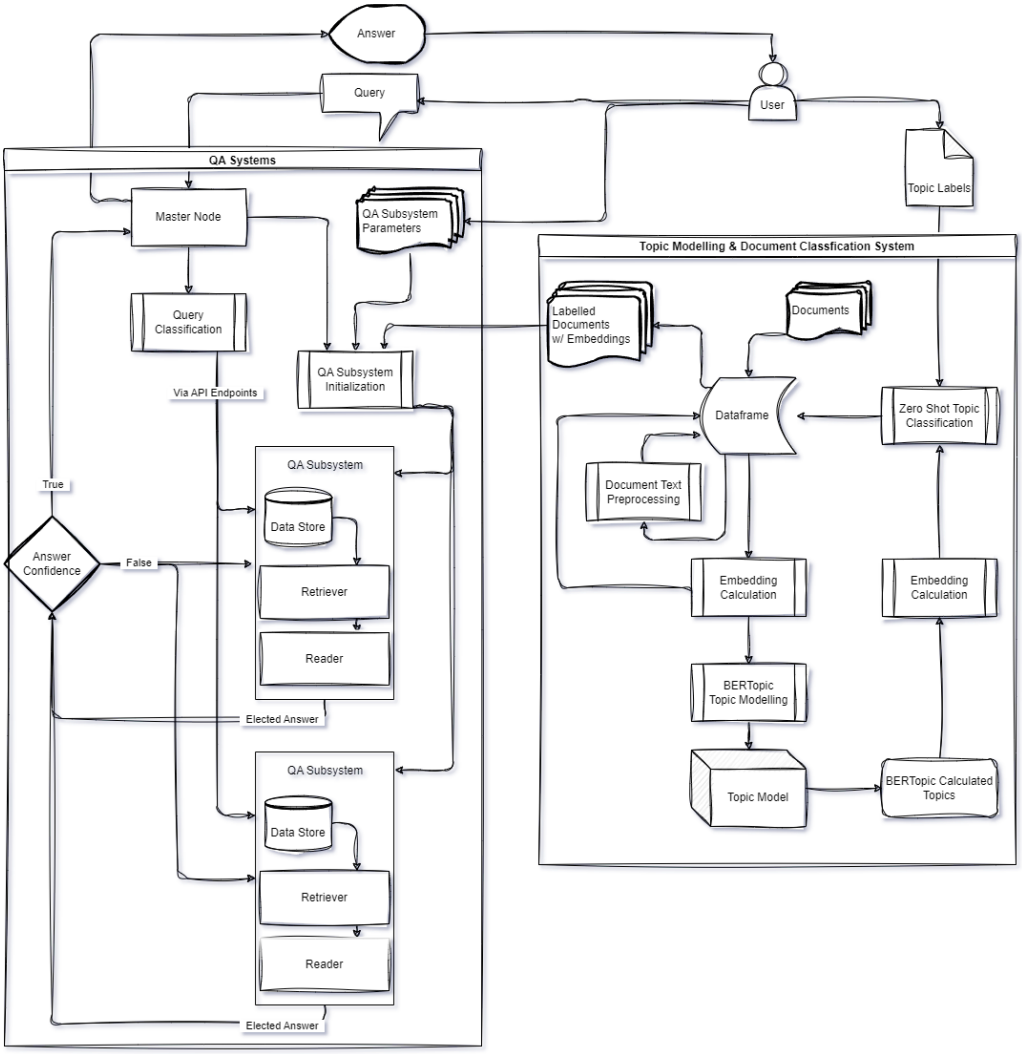
Σχήμα 4.5: Υπόδειγμα Αρχείου Διαμόρφωσης YAML

κάθε έγγραφο, μαζί και με τα embeddings του. Ύστερα μέσω ενός αρχείου Python θα δημιουργείται ένα λεξικό Python, σχήμα 4.4, το οποίο θα περιέχει όλες τις απαραίτητες παραμέτρους για να δημιουργηθεί ένα QA σύστημα. Το λεξικό αυτό θα αποθηκεύεται σε μορφή αρχείου YAML, όπως απεικονίζονται στο σχήμα 4.5, και θα χρησιμοποιείται ως αρχείο ρυθμιστικών πληροφοριών - configuration file από ένα αρχείο Python υπεύθυνο για την δημιουργία το QA συστήματος. Από την στιγμή που τα επιμέρους συστήματα QA μοιράζονται κοινή θεμελιώδη δομή η βασική διαφοροποίησή τους είναι τα αρχεία στα οποία ανατρέχουν και η ξεχωριστή περίσταση - instance του κάθε pipeline. Παράλληλα κάθε επιμέρους QA σύστημα θα διαθέτει ένα τερματικό σημείο API για να μπορεί να επικοινωνεί με τον βασικό κόμβο - master.

Ο βασικός κόμβος - master node (QA System Master Node στο σχήμα 4.3) του QA συστήματος είναι υπεύθυνος για την δρομολόγηση και ταξινόμηση των ερωτήσεων και την λήψη και αξιολόγηση των απαντήσεων. Αποτελεί τον κόμβο του συστήματος που με την βοήθεια της βιβλιοθήκης requests της Python αποστέλει και παραλαμβάνει τις ερωτήσεις και τις απαντήσεις αντίστοιχα. Εντός του Master Node λειτουργεί το query classification και παράδοση-παραλαβή των ερωτοαπαντήσεων, είναι δηλαδή ο κόμβος που επιτρέπει την αποκεντρωμένη φύση του συστήματος. Μόλις δεχτεί μια ερώτηση, αυτή θα ταξινομηθεί με τον Query Classifier και θα αποκτήσει ένα από τα ορισμένα topic labels, που αντιστοιχούν σε επιμέρους συστήματα QA, και ανάλογα με το αποτέλεσμα της ταξινόμησης και με την χρήση της βιβλιοθήκης FastAPI δημιουργείται ένα request για την πύλη - port στην οποία είναι καταχωρημένο το αντίστοιχο υποσύστημα QA. Αφού απαντηθεί η ερώτηση θα επιστραφεί στον Master Node η απάντηση, μέσω του API Endpoint. Είναι ταυτόχρονα υπεύ-

θυνος για την αξιολόγηση της εγκυρότητας των απαντήσεων που θα δεχτεί βάσει της μετρικής score που αποδίδει ο κόμβος του Reader του pipeline του εκάστοτε υποσυστήματος. Εάν η μετρική score, η οποία δέχεται τιμές μεταξύ του 0 και του 1, είναι υπό ενός ορισμένου κατωφλίου ορισμένο από τον χρήστη, υπάρχει η επιλογή να σταλέι η ερώτηση και στα λοιπά συστήματα, ώστε εάν έχει υπάρξει αστοχία στο classification των κειμένων να βρεθεί η απάντηση και αν κάμια από όλες τις δοσμένες απαντήσεις δεν έχει μετρική score, μεγαλύτερη του κατωφλίου, να μπορεί να επισημαίνεται ως ερώτηση χωρίς ξεκάθαρη απάντηση από το σύστημα.

4.7 Σχηματικη Απεικονιση Μεθοδολογιας



Σχήμα 4.6: Σχέδιο Λειτουργίας του Συστήματος

Παραπάνω στο σχήμα 4.6 περιγράφεται η συνολική ροή λειτουργίας του συνολικού συστήματος και η συνεργασία μεταξύ τους. Απεικονίζει την ταξινόμηση και την θεματική μοντελοποίηση των εγγράφων βάσει των προτιμήσεων του χρή-

στη, αλλά και την μεταβίβαση του συνόλου αυτών των πληροφοριών στο επόμενο στάδιο δημιουργίας του υποσυστήματος ερωτοαπαντήσεων. Παράλληλα περιγράφεται η συνεισφορά του χρήστη τόσο για την δημιουργία όσο και για την χρήση του καθολικού συστήματος.

5
Πειράματα - Αποτελέσματα

Σε αυτό το κεφάλαιο θα παρουσιαστούν και θα σχολιαστούν διάφορα πειράματα του πλήρους συστήματος της εργασίας. Θα γίνει μια επισκόπηση στην σημασία και την επίδραση που έχει η μεταβολή ορισμένων παραμέτρων του συστήματος, τόσο σε θεμελιώδη τμήματα της μεθοδολογίας όσο και σε μεταβολές και αλλαγές λεπτομεριών και παραμέτρων. Για τα πειράματα χρησιμοποιήθηκε το σύνολο δεδομένων - dataset "BBC News Dataset", καθώς εμπεριέχει κείμενα μεγάλου μήκους, πλούσιου σημασιολογικού περιεχομένου, ορισμένης θεματολογίας και παράλληλα είναι ήδη ταξινομημένα δεδομένα - labelled data, συνεπώς αξιοποιείται το γεγονός αυτό ως αναφορά για την αξιολόγηση ορισμένων λειτουργιών του συστήματος. Στο τμήμα της θεματικής μοντελοποίησης εξετάζεται κυρίως ο εύστοχος ή μη διαχωρισμός και η ομαδοποίηση των εγγράφων για την ταξινόμησή τους με την χρήση διαφόρων γνωστών και διαδεδομένων μοντέλων, όπως αυτά που αναφέρθηκαν παραπάνω, καθώς και μεταβολές που προκαλούν διάφορες τιμές παραμέτρων εντός της θεματικής μοντελοποίησης με την υποδομή του BERTopic. Δευτερευόντως, εξετάζονται και οι επιδόσεις του συστήματος σε κάθε επανάληψη. Στον τομέα του συστήματος ερωτοαπαντήσεων - QA System παρουσιάζεται η λειτουργία του ως πλήρες αυτόνομο - end-to-end σύστημα και η συνέργεια του διοργανωτικού κόμβου - master node με τα επιμέρους συστήματα. Παράλληλα γίνεται λόγος για τις επιδόσεις και απαιτήσεις πόρων του συστήματος και για την ευστοχία τόσο της ταξινόμησης των ερωτήσεων από τον master node όσο και για αυτή των απαντήσεων που επιστρέφει το σύστημα.

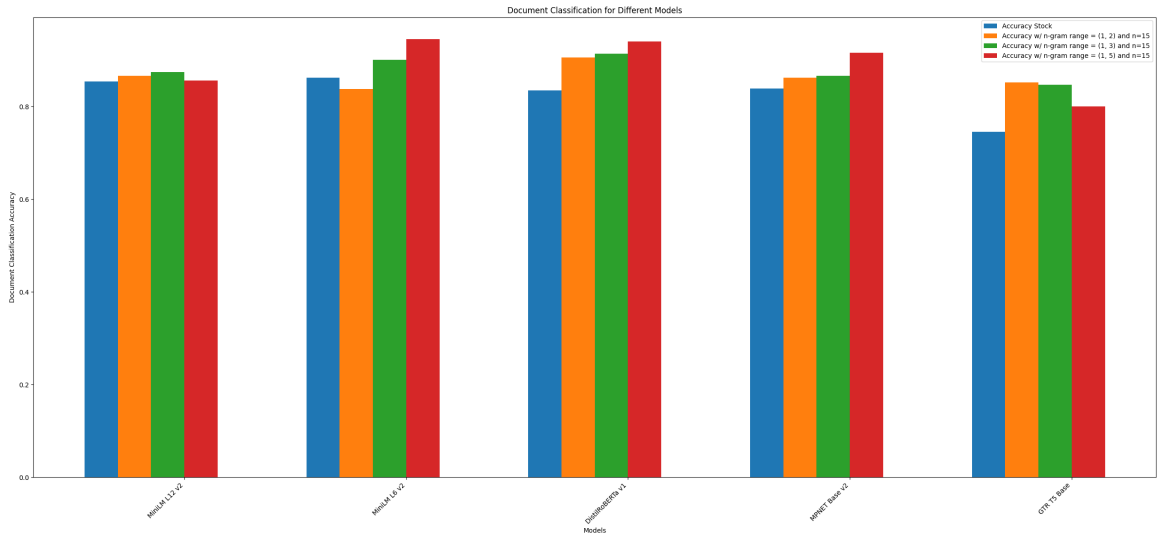
5.1 Αξιολογηση Ταξινομησης Εγγραφων

Για την θεματική μοντελοποίηση έγιναν πειράματα διαφόρων συνδυασμών καθώς και το τελικό επιθυμητό αποτέλεσμα στην προκειμένη περίπτωση δεν είναι πλήρως αντίστοιχο με αυτό της κύριας χρήσης του BERTopic. Συνεπώς υψηλή

σημασία για την αξιολόγηση των επιδόσεων ενός πειράματος έχει αφενώς μεν η σαφήνεια στην ομαδοποίηση των θεμάτων, αφετέρου δε, η ορθή απόδοση σημαντικών λέξεων σε κάθε θέμα ώστε να εκτελεστεί η ταξινόμηση των εγγράφων με την μέγιστη ευστοχία.

Τα πρώτα πειράματα με αυτό το dataset, διεξήχθησαν με μεθοδολογία αρκετά διαφορετική από την τρέχουσα στην εργασία, αίτιο αλλαγής της μεθοδολογίας αποτέλεσαν σαφώς τα μέτρια αποτελέσματα στο πρόβλημα της ταξινόμησης εγγράφων ανά θέματα

Τα αρχικά πειράματα της τρέχουσας μεθοδολογίας έγιναν με την χρήση των 5 προαναφερθέντων μοντέλων. ((α) "sentence-transformers/gtr-t5-base", (β) "all-mpnet-base-v2", (γ) "sentence-transformers/all-MiniLM-L6-v2", (δ) "sentence-transformers/all-MiniLM-L12-v2" και (ε) "sentence-transformers/all-distilroberta-v1") Τα μοντέλα ελήφθησαν από την ιστοσελίδα της Hugging Face και χρησιμοποιήθηκαν μέσω της βιβλιοθήκης SentenceTransformers για να υπολογιστούν τα embeddings των εγγράφων. Παράλληλα γίνεται χρήση της προεπιλεγμένης μεθόδου ομαδοποίησης HDBSCAN και με διάφορες τιμές για το εύρος n-grams. Η παράμετρος top_n_words παραμένει σταθερή και ίση με 15 για τα πειράματα πέρα από αυτή της προεπιλογής - default. Αυτό συμβαίνει για να υπάρχει κοινή είσοδος στον ταξινομητή των διανυσμάτων των θεμάτων, βάσει της βιβλιογραφίας του BERTopic το συνιστόμενο εύρος τιμών είναι από 10 έως 20, με το 20 να αποδεικνύεται χειρότερο στην ευστοχία της ταξινόμησης.



Σχήμα 5.1: Ευστοχία Ταξινόμησης θεμάτων για συνδυασμούς παραμέτρων BERTopic σε διάφορα μοντέλα

Όπως φαίνεται στο σχήμα 5.1, το μοντέλο MiniLM_L6_v2 είναι αυτό με τις βέλτιστες αποδόσεις και την μεγαλύτερη ευστοχία, γεγονός που εξυπηρετεί την μεθοδολογία του συστήματος καθώς υπολογίζει embeddings ομοίων διανυσματικών

διαστάσεων με το embeddigns που υπολογίζει το μοντέλο που θα χρησιμοποιεί ο κόμβος του Reader στα μετέπειτα στάδια. Επομένως τα πειράματα των υπόλοιπων παραμέτρων και του δεύτερου σκέλους γίνονται με το μοντέλο MiniLM_L6_v2. Σημειοταίον παράλληλα ότι αποδεικνύεται ότι στην συγκεκριμένη περίπτωση η απόσταξη γνώσης έχει λειτουργήσει εξαιρετικά, καθώς όχι μόνο δεν υστερεί από άλλα μεγαλύτερα μοντέλα, σε ορισμένες περιπτώσεις αποδίδει καλύτερα και γρηγορότερα και λόγω του γεγονότος ότι τα διανύσματα είναι διαστάσεων n=384 αντί για n=768 των υπολοίπων, τα τελικά αρχεία που εμπεριέχουν τα embeddings είναι το 63.7% του μεγέθους των αρχείων που παράγωνται από μοντέλα μεγαλύτερων διαστάσεων.

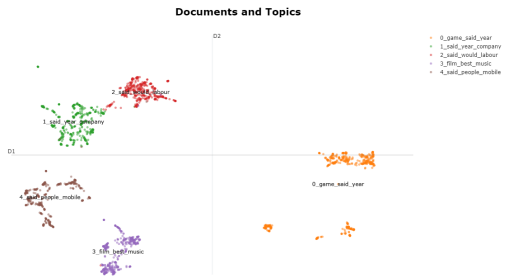
Παρακάτω (Πίνακας 5.1) δοκιμάζονται συνδυασμοί παραμέτρων για το μοντέλο MiniLM_L6_v2. Βασικές διαφορές και αλλαγές μεταξύ των παραμέτρων αποτελούν: (α) σύγκριση μεθόδων ομαδοποίηση k-Means και HDBSCAN, (β) αριθμός των ομάδων - clusters που δημιουργούνται ή ο ελάχιστος αριθμός εγγράφων ανά θέμα/ομάδα, (γ) το εύρος n-gram και ο αριθμός των πιο σημαντικών λέξεων ανά θέμα/ομάδα - topic/cluster.

Πίνακας 5.1: Αποτελέσματα ταξινόμησης θεμάτων ανά μοντέλο

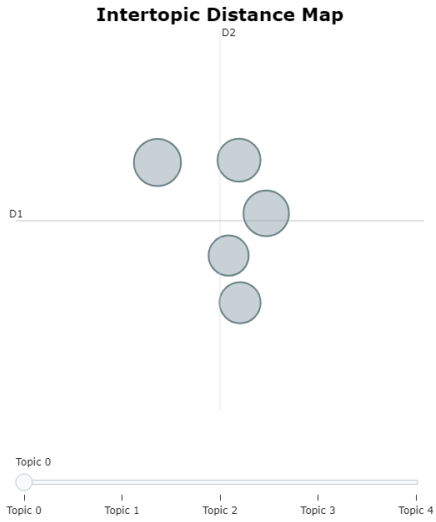
Μοντέλο	Clustering	No. of Clusters	N-gram	Top N	Accuracy
MiniLM L6 v2	k-Means	5	(1,1)	default	41.12%
		5	**(1,1)**	**15**	**94.43%**
		10	(1,1)	15	93.39%
		15	(1,1)	15	94.38%
		20	(1,1)	10	93.30%
		20	**(1,1)**	**15**	**94.74%**
		20	**(1,3)**	**15**	**95.78%**
		30	(1,1)	15	92.31%
		50	(1,1)	30	90.53%
	HDBSCAN	**default**	**(1,5)**	**15**	**94.56%**
		15	**(1,1)**	**15**	**91.69%**
MiniLM L12 v2	k-Means	**20**	**(1,1)**	**15**	94.02%
	HDBSCAN	default	(1,5)	15	85.62%

Από τον παραπάνω πίνακα 5.1 σημειώνονται 6 συνδυασμοί οι οποίοι έχουν αξιοσημείωτες επιδόσεις είτε κατά απόλυτη τιμή ή δεδομένων των παραμέτρων που χρησιμοποιήθηκαν και χρήζουν επιπλέουσας προσοχής. Συγκεκριμένα θα παρουσιαστεί οπτικά η ομαδοποίηση των εγγράφων και η διανομή των ομάδων στον χώρο. Οι 6 συνδυασμοί που επιλέγονται αποτελούνται από τους 3 με τις βέλτιστες επιδόσεις και ως επιπλέοντα σημεία αναφοράς 2 συνδυασμοί του μοντέλου "MiniLM L6 v2" με τις ελάχιστες υπολογιστικές απαιτήσεις, μικρό εύρος n-grams και χαμηλότερο αριθμών συνολικών ομάδων - clusters, καθώς και διαφορετικό αλγόριθμο ομαδοποίησης - clustering (k-Means και HDBSCAN), και ένας συνδυασμός με το μοντέλο "MiniLM L12 v2", το οποίο αποτελεί την μεγαλύτερη έκδοση του προηγουμένου και για αυτο επιλέγεται ο συνδυασμός με την βέλτιστη απόδοση.

Στα σχήματα 5.2 φαίνεται η ξεκάθαρη διαφοροποίηση που επιτυγχάνεται στην διαφοροποίηση των θεμάτων μεταξύ τους σε ευδιάκριτες συστάδες με αρκετά μι-



(α΄) Κατανομή εγγράφων ανά θέμα σε 2Δ χώρο



(β΄) Αποστάσεις μεταξύ ομάων/θεμάτων

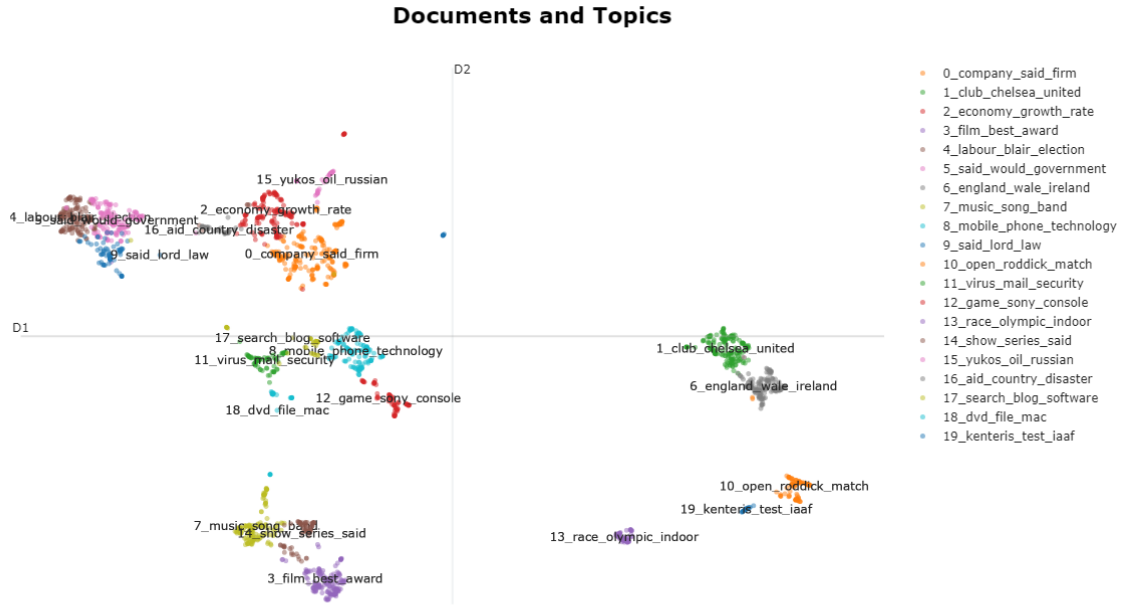
Σχήμα 5.2: Οπτικοποίηση MiniLM L6 v2, με k-Means: 5 clusters

κρό βαθμό επικάλυψης. Αυτό επιβεβαιώνεται σαφώς και με τα αποτελέσματα της ταξινόμησης και πιο συγκεκριμένα με τις μετρικές precision, recall και F1 score ανά κλάση, όπου όπως φαίνεται στο διάγραμμα, η ετικετά "sport" είναι σαφώς ορισμένη και σε μεγάλη απόσταση από τις υπόλοιπες ομάδες και επομένως πετυχαίνει precision score: 1.0, δεν δίνεται δηλαδή πουθενά αλλού η ετικέτα "sport".

Σχεδόν αντίστοιχα ευδιάκριτη διαφοροποίηση σε ομάδες ομάδων παρατηρείται και στις περιπτώσεις με αριθμό ομάδων ίσο με 20, σε όλους τους συνδυασμούς παραμέτρων του εύρους n-grams και μεγέθους μοντέλου, MiniLM_L6_v2 και MiniLM_L12_v2. Σε αυτές όμως τις περιπτώσεις υπάρχει μεγαλύτερη σύγχιση κυριώς στις ετικέτες 'business' και 'politics'.

Παρόμοια σύγχιση παρουσιάζουν και τα αποτελέσματα, των πειραμάτων που αξιοποιήσαν HDBSCAN για την ομαδοποίηση. Χαρακτηριστικό του όμως ήταν το γεγονός ότι δημιουργείται ένα, ως θέμα/ομάδα για τις ακραίες ή 'αταίριαστες' περιπτώσεις - outliers, για αυτό το λόγω η αρίθμηση των θεμάτων ξεκινάει από το -1. Έτσι λοιπόν εξηγείται και η μικρή έλλειψη των πειραμάτων με αξιοποίηση του HDBSCAN καθώς όσα αρχεία ενταχθούν στο topic αυτό θα έχουν πολύ λιγότερες πιθανότητες να ταξινομηθούν σωστά στην συνέχεια βάσει των σημαντικότερων λέξεων της ομάδας.

Ο παρακάτω πίνακας 5.2 παρουσιάζει τα F1-score ανά κλάση για κάθε μια από τις ετικέτες θεμάτων στην ταξινόμηση μηδενικής βολής - zero shot classification. Παρατηρείται ότι οι συνδυασμοί που χρησιμοποιούν το μοντέλο MiniLM L6 v2 και τον k-Means για αλγόριθμο ομαδοποίησης παρουσιάζουν τα καλύτερα αποτελέσματα στις αναλυτικές μετρικές αξιολόγησης ταξινόμησης. Άξιο αναφοράς είναι το γεγονός ότι η επιδόσεις του πιο βασικού και μικρού μοντέλου/συνδυασμού (MiniLM_L6_v2 με 5 ομάδες και (1,1) εύρος n-gram) και είναι τοσό κοντά σε αυτές των υπολοίπων παρά τις διαφορές σε μέγεθος δεδομένων, την περιπλοκότητα των tokenizers, στην



Σχήμα 5.3: Κατανομή εγγράφων ανά θέμα στον 2Δ χώρο, k-Means: 20 Clusters

	Topic	Count	Name	Representation	Representative_Docs
0	-1	262	-1_said_year_company_market	[said, year, company, market, bank, share, sal...	[bank england left interest rate hold widely p...
1	0	415	0_said_labour_party_election	[said, labour, party, election, would, governm...	[michael howard finally revealed full scale pl...
2	1	344	1_mobile_people_phone_said	[mobile, people, phone, said, game, technology...	[mobile phone still enjoying boom time sale ac...
3	2	337	2_england_game_club_player	[england, game, club, player, side, team, wale...	[sale shark director rugby philippe saint andr...
4	3	172	3_film_best_award_actor	[film, best, award, actor, oscar, star, direct...	[aviator named best film golden globe award st...
5	4	118	4_band_music_album_song	[band, music, album, song, best, award, year, ...	[three prestigious grammy award hit vertigo st...
6	5	88	5_open_roddick_match_seed	[open, roddick, match, seed, set, final, austr...	[top seed lindsay davenport booked place last ...
7	6	82	6_economy_growth_rate_economic	[economy, growth, rate, economic, dollar, year...	[created fewer job expected december analyst s...
8	7	70	7_race_olympic_indoor_world	[race, olympic, indoor, world, championship, c...	[athletics fan endured year mixed emotion stun...
9	8	49	8_show_series_bbc_channel	[show, series, bbc, channel, comedy, said, cel...	[bbc flagship pop music programme top pop move...
10	9	36	9_ebbers_fraud_firm_worldcom	[ebbers, fraud, firm, worldcom, sec, company, ...	[former worldcom chief bernie ebbers denied cl...
11	10	32	10_airline_boeing_flight_air	[airline, boeing, flight, air, cost, passenger...	[airline attendant suspended inappropriate ima...
12	11	31	11_yukos_russian_gazprom_russia	[yukos, russian, gazprom, russia, oil, rosneft...	[russia president defended purchase yukos key ...
13	12	28	12_kenteris_test_iaaf_doping	[kenteris, test, iaaf, doping, thanou, drug, c...	[greek sprinter kostas kenteris katerina thano...
14	13	28	13_music_player_digital_chart	[music, player, digital, chart, ipod, industry...	[music downloading rejected free peer peer ser...
15	14	27	14_car_fiat_sale_bmw	[car, fiat, sale, bmw, motor, vehicle, year, n...	[car firm general motor ford forced cut produc...
16	15	26	15_dvd_file_system_technology	[dvd, file, system, technology, peer, network,...	[peer peer network stay verge exploited commer...
17	16	23	16_profit_sale_share_year	[profit, sale, share, year, business, said, re...	[behemoth general electric posted jump quarter...
18	17	22	17_sri_lanka_disaster_indonesia	[sri, lanka, disaster, indonesia, tsunami, aff...	[indonesian indian hong kong stock market reac...
19	18	20	18_oil_crude_price_barrel	[oil, crude, price, barrel, cairn, production,...	[oil price fallen heavily second day closing t...
20	19	15	19_china_embargo_dam_project	[china, embargo, dam, project, straw, arm, lif...	[embargo arm export china likely lifted next s...

Σχήμα 5.4: Στοιχεία ομάδων/θεμάτων για ομαδοποίηση με HDBSCAN

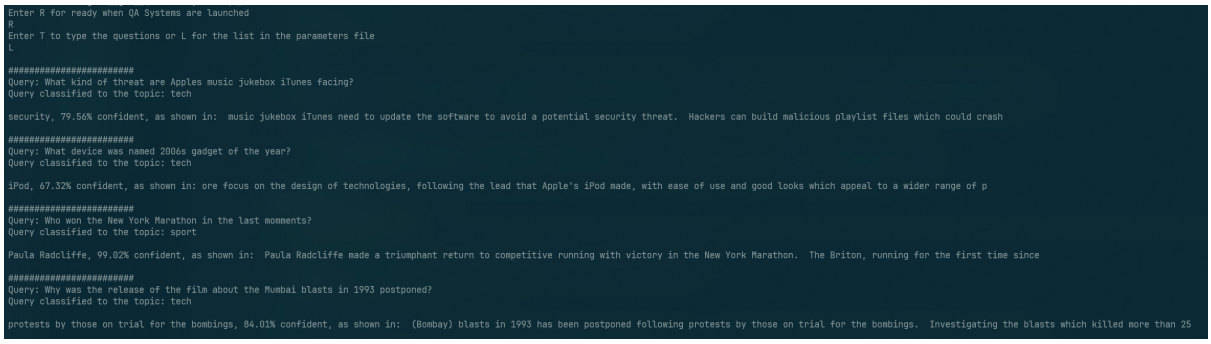
περίπτωση του μεγαλυτερου έυρους n_gram και τον αισθητά χαμηλότερο χρόνο εκτέλεσης, n-gram = (1,1) : 13s και n-gram = (1,3) : 24s, σχεδόν διπλάσια χρονική απαίτηση για τον υπολογισμό και την παραγωγή του θεματικού μοντέλου.

Πίνακας 5.2: F1-Score ανά κλάση για συνδυασμούς παραμέτρων BERTopic

Μοντέλο	Clustering	Clusters	N-gram	Top N	Topic	F1-Score
MiniLM L6 v2	k-Means	5	(1,1)	15	Business	0.908
					Entertainment	0.932
					Politics	0.943
					Sport	0.999
					Tech	0.934
		20	(1,1)	15	Business	0.911
					Entertainment	0.938
					Politics	0.951
					Sport	0.998
					Tech	0.935
		20	(1,3)	15	Business	0.893
					Entertainment	0.944
					Politics	0.935
					Sport	0.998
					Tech	0.896
	HDBSCAN	default	(1,5)	15	Business	0.814
					Entertainment	0.906
					Politics	0.711
					Sport	0.980
					Tech	0.782
		15	(1,1)	15	Business	0.815
					Entertainment	0.907
					Politics	0.929
					Sport	0.999
					Tech	0.946
MiniLM L12 v2	k-Means	20	(1,1)	15	Business	0.814
					Entertainment	0.95
					Politics	0.915
					Sport	0.997
					Tech	0.95

5.2 Πειραματα Συστηματοσ Ερωτοαπαντησεων

Οι δοκιμές και τα πειράματα αυτού του σκέλους του συστήματος της διπλωματικής εργασίας βασίζονται στην εύρυθμη εκκίνηση και λειτουργία του υποσυστήματος διαχείρισης και των επιμέρους υποσυστημάτων ερωτοαπαντήσεων, καθώς και την αλληλεπίδραση αυτών με τον χρήστη. Υπολογίζονται παράλληλα μετρικές για τις ευστοχίες του συστήματος κατά την ταξινόμιση των ερωτήσεων και την επιτυχή δρομολόγησή τους στο κατάλληλο υποσύστημα. Γίνεται έλεγχος των τερματικών σημείων - endpoints των REST APIs μέσα από την διεπαφή τους από ένα σύστημα περιήγησης. Δοκιμάζεται φυσικά και η ικανότητα του συστήματος να απαντάει σωστά σε ερωτήσεις πάνω στα δεδομένα του dataset.



Σχήμα 5.5: Αλληλεπιδραση με το QA σύστημα μέσω Command Line Interface

Με την εκκίνηση της εφαρμογής του διοργανωτή κόμβου - master node, εκκινούν ως ξεχωριστές διεργασίες τα επιμέρους συστήματα και τρέχουν ως εφαρμογή του FastAPI στην προκαθορισμένη από τα YAML configuration files διεύθυνση. Αφότου εγκαθιδρυθεί η σύνδεση των επιμέρους εφαρμογών μπορεί το σύστημα να δεχτεί ερωτήσεις είτε από τον χρήστη ή από προκαθορισμένη λίστα ερωτήσεων.

Στις εικόνες 5.5 φαίνεται η CLI διεπαφή του συστήματος και 5.6 η διεπαφή του REST API όσο αυτό βρίσκεται σε λειτουργία και ένα πείραμα της συνάρτησης POST στο τερματικό σημείο από την εφαρμογή περιήγησης. Στο προκειμένο πείραμα έχουν δημιουργηθεί 5 ξεχωριστά επιμέρους υποσυστήματα ερωτοαπαντήσεων, 1 για κάθε ετικέτα του συνόλου δεδομένων.

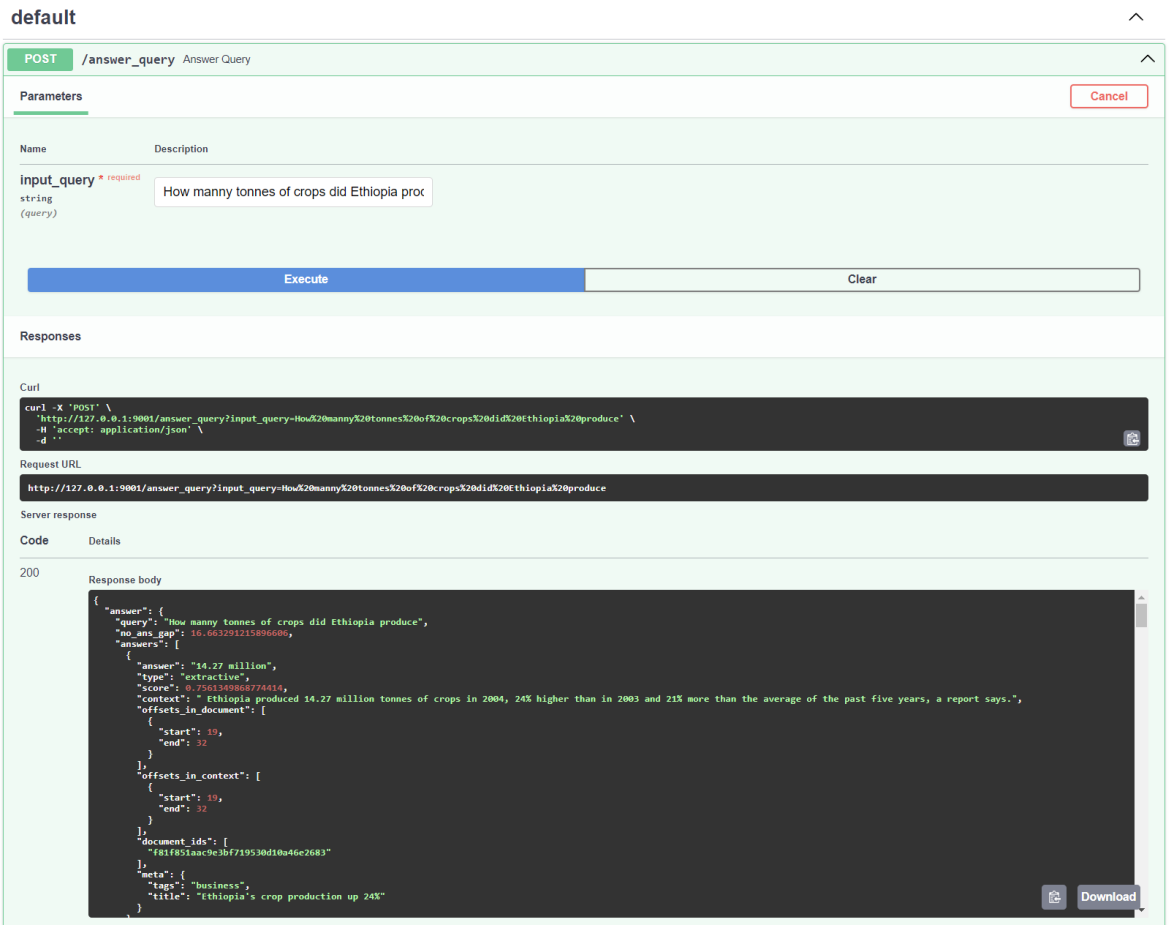
5.2.1 Αξιολόγηση Ταξινομητή Ερωτήσεων - Query Classifier

Για τον κόμβο του Query Classifier χρησιμοποιήθηκαν 2 μοντέλα προτεινόμενα για ταξινόμηση μηδενικής βολής: (α) "facebook/bart-large-mnli" και (β) "cross-encoder/nli-deberta-base". Σε αυτά χρησιμοποιήθηκαν διάφορα σύνολα ερωτήσεων: (α) 50 ερωτήσεις χειρόγραφες και στοχευμένες ειδικά για συγκεκριμένα τμήματα κειμένων του dataset, αρκετές εκ των οποίων είχαν διττό χαρακτήρα και λέξεις που θα μπορούσαν να ανήκουν σε άλλη θεματολογία (β) 50 γενικές ερωτήσεις πάνω στις εν λόγω θεματολογίες και (γ) 200 ερωτήσεις παρηγμένες από το ChatGPT αναφερόμενες στο dataset που χρησιμοποιείται.

Πίνακας 5.3: Αποτελέσματα ταξινόμησης ερωτήσεων με Zero Shot Classification

Μοντέλο	Ερωτήσεις (α)	Ερωτήσεις (β)	Ερωτήσεις (γ)
facebook/bart-large-mnli	90%	88%	93.5%
cross-encoder/nli-deberta-base	86%	84%	88%

Από τον παραπάνω πίνακα 5.3 επομένως φαίνεται ότι το μοντέλο ταξινόμηση μηδενικής βολής "facebook/bart-large-mnli" αποδίδει καλύτερα και είναι όμοιο με αυτό που χρησιμοποιήθηκε και στο προηγούμενο σκέλος της εργασίας. Επίσης παρατηρείται ότι οι ερωτήσεις που αναφέρονται συγκεκριμένα στο dataset και προέρχονται από μηχανική παραγωγή φυσικής γλώσσας είναι αυτές που ταξινομούνται



Σχήμα 5.6: Response του REST API στο POST request, δοκιμή του API από μηχανή περιήγησης

καλύτερα, αλλά και οι ερωτήσεις τύπου (α) παρά τις πολυπλοκότητές τους ταξινομούνται σωστά σε μεγάλο βαθμό.

5.2.2 Εμπειρική Αξιολόγηση Ευστοχίας των Απαντήσεων του Συστήματος

Οι παραπάνω ερωτήσεις τύπου (α) είναι ερωτήσεις απαντήσεις των οποίων εντοπίζονται σε συγκεκριμένα σημεία των κειμένων του dataset. Οι ερωτήσεις αυτές δόθηκαν διαδοχικά ως είσοδος στο σύστημα και αξιολογήθηκαν τα αποτελέσματα, βάσει του σωστού εντοπισμού της απάντησης. Από τις 50 ερωτήσεις: (α) οι 5 ταξινομήθηκαν λάθος, αλλά απαντήθηκαν όταν ερωτήθηκαν τα υπόλοιπα συστήματα, (β) οι 2 απαντήθηκαν πάλι στην δεύτερη φάση των απαντήσεων, λόγω εσφαλμένης ταξινόμησης των εγγράφων, (γ) οι 7 απαντήθηκαν λάθος παρά την ορθή ταξινόμησή τους και (δ) οι 36 απαντήθηκαν σωστά εξ' αρχής με σωστή ταξινόμηση της ερώτησης και των εγγράφων. Συνεπώς αν γενικευθεί το πείραμα αυτό υπήρξε 72% πλήρης ευστοχία, 86% ευστοχία στις απαντήσεις

5.2.3 Αξιολόγηση χρόνου εκτέλεσης

Συγκρίθηκε επίσης ο χρόνος που χρειάστηκε για να απαντηθεί μια ερώτηση μέσω ενός συστήματος αυτής της φιλοσοφίας με αυτόν ενός συστήματος ομογενές και μη διαιρεμένο 5.4. Ερευνάται δηλαδή ποια είναι η χρονική ποινή της επανειλημμένης ταξινόμησης των ερωτήσεων και αν αυτή είναι ανάλογη της χρονικής ποινής του σαφώς μεγαλύτερου αριθμού εγγράφων στο ενιαίο Document Store, στα οποία θα πρέπει να ανατρέξει ο Retriever. Σημειοταίο ότι δεν υπολογίζεται κάποιος χρόνος ποινής για την αλληλεπίδραση με τα τερματικά σημεία του API, καθώς εάν δεν είναι όλα τα έγγραφα και τα αρχεία αποθηκευμένα τοπικά θα εμφανίζονται επιπλέοντες χρόνοι και στις δύο περιπτώσεις. Το πείραμα έγινε με τις ίδιες ακριβώς ερωτήσεις, όμως στην πρώτη περίπτωση το pipeline έτρεχε την διεργασία του διαδοχικά, ενώ στην δεύτερη περίπτωση πρώτα γινόταν η ταξινόμηση της ερώτησης και μετά έτρεχε η διεργασία του επιλεγμένου pipeline. Επίσης προστέθηκαν προσομοιώσεις για διάφορες τιμές ευστοχίας της ταξινόμησης των ερωτήσεων.

Προσέγγιση	Classification Accuracy					
	100%		95%		90%	
	Seconds per query					
Ενιαίο Σύστημα	0.557	Δ=-7%	0.557	Δ=0.18%	0.557	Δ=1%
Domain Specific	0.52		0.558		0.563	

Πίνακας 5.4: Απαιτούμενος χρόνος για απάντηση ερώτησης μεταξύ των 2 τύπων συστημάτων

Επομένως αποδεικνύεται εν ονλίγοις ότι οι επιπλέουσες υπολογιστικές απαιτήσεις αμφότερων προσεγγίσεων σχεδόν εξουδετερόνονται από άποψη απαιτούμενου χρόνου εκτέλεσης. Στον πίνακα 5.4 υπολογίζεται κατά μέσο όρο ο απαιτούμενος χρόνος για την απάντηση μίας ερώτησης και η ποσοστιαία διαφορά μεταξύ των 2 τύπων συστημάτων.

6 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στο παρόν κεφάλαιο θα γίνει μια σύντομη ανασκόπηση και αναδρομή του συστήματος της διπλωματικής εργασίας μαζί με συμπεράσματα που αντλούνται από την έκβαση των δοκιμών και πειραμάτων. Εν συνεχεία, γίνεται αναφορά σε πιθανές μελλοντικές επεκτάσεις.

6.1 Σύνοψη

Συνοψίζοντας, στο πλαίσιο της διπλωματικής εργασίας αναπτύχθηκε και υλοποιήθηκε ένα σύστημα οργάνωσης και διαχείρισης εγγράφων για αποκεντρωμένα συστήματα ερωτοαπαντήσεων ορισμένου θέματος. Χρησιμοποιείται η υποδομή BERTopic σε συνδυασμό με μοντέλα μετασχηματιστών προτάσεων ώστε να αναπαρασταθούν διανυσμαικά τα έγγραφα. Με αλγόριθμους c-TF-IDF και αλγόριθμους ομαδοποίησης δημιουργούνται θεματικά μοντέλα μέσω των οποίων ταξινομούνται τα έγγραφα σε θέματα επιλεγμένα από τον χρήστη, με την χρήση zero shot classification. Τα ταξινομημένα αρχεία ύστερα αποθηκεύονται σε Document Stores και ανάλογα με τις θεματολογίες που έχει επιλέξει ο χρήστης, δημιουργούνται συστήματα ερωτοαπαντήσεων ορισμένου θέματος - domain-specific question answering systems, τα οποία ελέγχονται και διαχειρίζονται από έναν διοργανωτικό κόμβο - master node. Ύστερα οι ερωτήσεις θέτονται από τον χρήστη ως είσοδος στον master node ταξινομούνται στις υποψήφιες ετικέτες θεμάτων και μεταβιβάζονται στο αντίστοιχο επιμέρους σύστημα ερωτοαπαντήσεων, το οποίο ανατρέχει στα κείμενα των εγγράφων και επιστρέφει την πιο ταιριαστή, στην ερώτηση, απάντηση. Αποτέλεσμα είναι ένα πλήρες σύστημα με σαφώς μειωμένες υπολογιστικές απαιτήσεις και παρόμοιες επιδόσεις από άποψη χρόνου εκτέλεσης και ευστοχίας απαντήσεων καθώς και με την απαραίτητη πληροφορία αναδιοργάνωσης και παραμετροποίησης.

6.2 Γενικα Συμπερασματα

- Η θεματική μοντελοποίηση - topic modelling των εγγράφων μπορεί μεταξύ άλλων να αποτελέσει χρήσιμο εργαλείο για την σημασιολογική ταξινόμηση τους. Ο συνδυασμός μεθοδολογιών και τεχνικών ταξινόμησης όπως οι πιο σημαντικές λέξεις για κάθε ομάδα ενός θεματικού μοντέλου και η ταξινόμηση μηδενικής βολής - zero shot classification επιστρέφουν πολύ αξιόλογα αποτελέσματα με υψηλές τιμές σε διάφορες μετρικές αξιολόγησης και δίνουν παράλληλα την δυνατότητα να διατηρηθούν οι διανυσματικές αναπαραστάσεις - embeddings των κειμένων για μετέπειτα διεργασίες και ταυτόχρονα να γίνουν οι σχεδόν ελάχιστες συγκρίσεις για την ταξινόμηση των εγγράφων.
- Η εργασία αποτελεί έως έναν βαθμό proof-of-concept για μια διαφορετική προσέγγιση συστημάτων ερωτοαπαντήσεων, όπου αξιοποιούνται οι ευρύτερες ιδέες των αποκεντρωμένων συστημάτων σε συνδυασμό εκτενούς αυτοματοποιημένου text labelling and classification. Αποδεικνύεται ότι δεν η αρχιτεκτονική και μεθοδολογία ενός τέτοιου συστήματος όχι μονο δεν είναι αποτρεπτική, αλλά υπό ορισμένες συνθήκες είναι και πιο αποτελεσματική. Μειώνονται οι απαιτήσεις από την μνήμη, τόσο για την αποθήκευση όσο και για την εκτέλεση, αυξάνονται σημαντικά και καθοριστικά οι δυνατότητες για παραμετροποίηση και personalization του συστήματος καθώς "ανακυκλώνονται" και επαναχρησιμοποιούνται οι πληροφορίες που το επιτρέπουν, BERTopic topics και embeddings. Ταυτόχρονα προσφέρει έναν επιπλέοντα βαθμό ασφάλειας στο σύστημα σε περίπτωση που είναι πλήρως αποκεντρωμένο και κάθε υποσύστημα είναι σε διαφορετικό μηχάνημα ή και χώρο, οπότε με την εξαίρεση του κυρίου κόμβου - master node, όλα τα υπόλοιπα υποσυστήματα είναι πλήρως ανεξάρτητα το ένα από το άλλο και μπορούν να λειτουργούν ακόμα και αν κάποιο πάψει. Επιπρόσθετα φάνηκε ότι η αρνητική επίδραση στις συνολικές επιδόσεις είναι αρκετά μικρή και εύκολη να αντιμετωπιστεί, παρά την επιπλέουσα πολυπλοκότητα του συστήματος. Επομένως τα παραπάνω καθιστούν αυτή την προσέγγιση για ένα αντίστοιχο σύστημα, έχουσα πρακτικής αξίας και ενδιαφέροντος.

6.3 Μελλοντικες Επεκτασεις & Εργασια

Τέλος παρουσιάζονται προτάσεις και δυνητικές επεκτάσεις της εργασίας αυτής για την εξέλιξη ή βελτίωση ενός αντίστοιχου συστήματος ή της ευρύτερης φιλοσοφίας και προσέγγισης.

Πρώτη εξέλιξη του συστήματος θα μπορούσε να είναι η πλήρης παραλληλοποίηση των διεργασιών όπου αυτό είναι δυνατό. Ήδη σε μεγάλο βαθμό οι υπολογισμοί και οι δοκιμές γίνονται με χρήση πολυ-πύρηνων επεξεργαστικών μονάδων, όπως GPUs, κυρίως για τους Readers και τους υπολογισμούς των embeddings. Παρ'όλα αυτά η αποκεντρωμένη φύση του συστήματος ενδεχομένως να επιτρέπει την πρόβλεψη και προετοιμασία του συστήματος για ορισμένες προσεχής διεργασίες αλλά και η εκτέλεση αυτών ταυτόχρονα με σκοπό την χρονική βελτιστοποίηση.

Μια ενδιαφέρουσα κατεύθυνση και εξέλιξη του συστήματος θα ήταν η ενσωμάτωση διαλογικού χαρακτήρα και μνήμης, δηλαδή παρόμοια με μια από τις πρώτες και θεμελειώδεις αλλαγές που υπέστει η υπηρεσία του ChatGPT της OpenAI, να έχει την δυνατότητα ο χρήστης, να θυμάται το σύστημα το αποτέλεσμα της πρώτης ή μιας προηγούμενης ταξινόμησης και αντί να ταξινομείται κάθε μια από τις ερωτήσεις να "θυμάται" το σύστημα ότι έχει ξεκινήσει ένας διάλογος σε κάποια συγκεκριμένη θεματολογία και θα πάψει όταν το επιλέξει ο χρήστης. Η λειτουργία του ChatGPT επιτρέπει στον χρήστη να επιστρέψει σε μια παλαιότερη συζήτηση που είχε με το σύστημα και αυτή να συνεχιστεί από εκεί που έμεινε. Με αυτόν τον τρόπο δεν χρειάζεται να ξαναπεριγράψει κανείς τις περιβάλλουσες συνθήκες και αν επιθυμει κανείς να ξεκινήσει εκ νέου επιλέγει την επιλογή νέας συνομιλίας. Έτσι θα μπορεί να παρακάμπτεται ένα βήμα της μεθοδολογίας και να γίνεται όλη η διαδικασία αισθητά γρηγορότερη και πιο αποτελεσματική.

Άλλη επέκταση μπορεί να αποτελέσει η εξερεύνηση περισσότερων επιλογών προσωπικοποίησης - personalization του συστήματος. Να μπορεί δηλαδη ο χρήστης να επιλέγει όχι μόνο τα θέματα στα οποία θα χωριστούν τα έγγραφα αλλά και μια πιο σύθνετη σύνθεση και συνδυασμό θεματολογία και εγγράφων για κάθε επιμέρους υποσύστημα. Επέκταση που δεν απαιτεί σοβαρές αλλαγές στον κορμό του συστήματος, παρά μόνο πολλές και συγκεκριμένες πληροφορίες για την φύση του προβλήματος που δύναται να επιλύσει.

Τέλος μια παραλλαγή ή εξέλιξη θα μπορούσε είναι η ενσωμάτωση του παραγωγικού στοιχείου της φυσικής γλώσσας, να χρησιμοποιηθούν δηλαδή μοντέλα παραγωγής φυσικής γλώσσας σε συνδυασμό με τους αλγορίθμους εύρεσης και ανάκτησης πληροφορίας, ώστε να απαντώνται οι ερωτήσεις με επαυξημένη παραγωγή με ανάκτηση - retrieval augmented generation - RAG και να μοιάζουν με απαντήσεις που θα έδινε ένας άνθρωπος.

Βιβλιογραφία

[1] J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Ábrego, J. Ma, V. Y. Zhao, Y. Luan, K. B. Hall, M.-W. Chang, and Y. Yang, "Large dual encoders are generalizable retrievers," 2021.

[2] E. D. Liddy, "Natural language processing.," 2001.

[3] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. The mathematical theory of communication., Champaign, IL, US: University of Illinois Press, 1949.

[4] N. Chomsky, *Syntactic Structures*. De Gruyter, Dec. 1957.

[5] *Language and Machines*. National Academies Press, Jan. 1966.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[7] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[8] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[10] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[12] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," 2019.

[13] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[14] G. Luo, X. Huang, C.-Y. Lin, and Z. Nie, "Joint entity recognition and disambiguation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (L. Màrquez, C. Callison-Burch, and J. Su, eds.), (Lisbon, Portugal), pp. 879–888, Association for Computational Linguistics, Sept. 2015.

[15] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," *IEEE Transactions on Multimedia*, p. 1–1, 2015.

[16] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[18] D. Angelov, "Top2vec: Distributed representations of topics," 2020.

[19] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in Sociology*, vol. 7, 2022.

[20] B. Das and S. Chakraborty, "An improved text sentiment classification model using tf-idf and next word negation," 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[22] R. C. Staudemeyer and E. R. Morris, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," 2019.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," 2020.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[26] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019.

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[29] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," 2021.

[30] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, p. 1608, Sep 2016.

[31] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," 2016.

[32] D. N. Nastos, *Design and Development of Greek Open-Domain Question Answering System*. Διπλωματική Εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2022.

[33] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, "Dense passage retrieval for open-domain question answering," 2020.

[34] W. McKinney *et al*., "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, Austin, TX, 2010.

[35] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[37] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[38] X. Jin and J. Han, *K-Means Clustering*, pp. 563–564. Boston, MA: Springer US, 2010.

[39] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining* (J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds.), (Berlin, Heidelberg), pp. 160–172, Springer Berlin Heidelberg, 2013.

[40] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, p. 333–389, 2009.

[41] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question answering systems: Survey and trends," *Procedia Computer Science*, vol. 73, pp. 366–375, 2015. International Conference on Advanced Wireless Information and Communication Technologies (AWICT 2015).

[42] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[43] A. Alcoforado, T. P. Ferraz, R. Gerber, E. Bustos, A. S. Oliveira, B. M. Veloso, F. L. Siqueira, and A. H. R. Costa, "ZeroBERTo: Leveraging zero-shot text classification by topic modeling," in *Lecture Notes in Computer Science*, pp. 125–136, Springer International Publishing, 2022.

[44] S. Velu, "An empirical science research on bioinformatics in machine learning," *JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES*, vol. spl7, 02 2020.

[45] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," 2021.

[46] Z. Zeng, Y. Li, J. Yong, X. Tao, and V. Liu, "Multi-aspect attentive text representations for simple question answering over knowledge base," *Natural Language Processing Journal*, vol. 5, p. 100035, 2023.

[47] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, pp. 3713–3744, Jan 2023.

[48] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[49] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 2222–2232, oct 2017.

[50] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly," 2020.