



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

Ανάπτυξη Δυναμικού Συστήματος Ερωταπαντήσεων με Πηγή το Διαδίκτυο

Σκαπέτης Χρήστος
ΑΕΜ: 9378

Επιβλέποντες: **Ανδρέας Α. Συμεωνίδης**
Καθηγητής Α.Π.Θ.

Νικόλαος Μάλαμας
Υποψήφιος Διδάκτορας

7 Νοεμβρίου 2023

Περίληψη

Η Τεχνητή Νοημοσύνη αποτελεί έναν τομέα της τεχνολογίας που καθημερινά πραγματοποιεί άλματα. Αποτελεί την αιχμή του δόρατος στην επίλυση πληθώρας προβλημάτων αλλά και στην παροχή υπηρεσιών υψηλού επιπέδου. Οι εμπορικές εφαρμογές, αν και με απεριόριστες προοπτικές, τα τελευταία χρόνια εστιάζουν κυρίως στη Συνομιλητική Τεχνητή Νοημοσύνη ή στην Τεχνητή Νοημοσύνη Επεξεργασίας Βίντεο και Εικόνας, με αποτέλεσμα τον ενθουσιασμό της αγοράς και των χρηστών, οι οποίοι αποκτούν όλο και μεγαλύτερες απαιτήσεις, ως προς την ποιότητα και τις δυνατότητες των εργαλείων τους.

Οι συνομιλητικοί βοηθοί, Chatbots, έχουν διεισδύσει στην καθημερινότητα, με πληθώρα μορφών και χρήσεων. Η παραδοσιακή μορφή τους, που απαντούσε ένα κλειστό σύνολο απλοϊκών ερωτήσεων, αντικαταστάθηκε από μοντέρνα συστήματα και εφαρμογές που προτείνουν, συζητούν, συνομιλούν, διασκεδάζουν, συμβουλεύουν. Είτε ως εφαρμογές στο κινητό του χρήστη, είτε διαθέσιμοι online, οι σύγχρονοι βοηθοί αποτελούν πλέον κομμάτι της καθημερινής ρουτίνας και των εργαλείων που την καθιστούν πιο ανεκτή και ευχάριστη. Αυτή η καθολική πλέον παρουσία τους καθιστά αναγκαίο τον εκλεκτισμό αυτών των υπηρεσιών προσθέτοντας νέα στοιχεία και παραμέτρους, ανάλογα με την εξειδίκευση κάθε χρήστη και των αναγκών του.

Αυτές οι εφαρμογές, όντας συνεχώς συνδεδεμένες στο διαδίκτυο, θυσιάζουν την ιδιωτικότητα του χρήστη, καθιστώντας τον επιρρεπή σε κακόβουλο λογισμικό ή απάτες. Παράλληλα, όμως, η πλειοψηφία αυτών των εφαρμογών αδυνατούν να χρησιμοποιήσουν τον μεγαλύτερο όγκο της πληροφορίας η οποία είναι διαθέσιμη στο διαδίκτυο, παρότι είναι συνδεδεμένες σε αυτό. Τέλος, τα υπάρχοντα συστήματα δεν έχουν προσωποποιημένο χαρακτήρα. Οι απαντήσεις και οι υπηρεσίες τους σπανίως προσαρμόζονται στον εκάστοτε χρήστη.

Σε αυτήν τη διπλωματική εργασία γίνεται μια προσπάθεια ανάπτυξης ενός ψηφιακού βοηθού ερωταπαντήσεων που αναζητά δυναμικά τις απαντήσεις στον ιστό. Αυτό το σύστημα, χρησιμοποιώντας προεκπαιδευμένα μοντέλα αλλά και τεχνολογίες λογισμικού, απαντάει τις ερωτήσεις του χρήστη αξιοποιώντας το σύνολο της online πληροφορίας. Η φορητότητα του συστήματος μπορεί να εξασφαλίσει μια ιδιωτική εμπειρία καθώς και απαντήσεις από εξειδικευμένες πηγές δεδομένων.

Τα πειραματικά αποτελέσματα αποδεικνύουν ότι το συγκεκριμένο σύστημα είναι επαρκώς ακριβές, ενώ ταυτόχρονα επιτυγχάνει τους σκοπούς για τους οποίους αναπτύχθηκε. Θυσιάζεται η αναλυτικότητα και το συνομιλητικό ύφος των συμβατικών ψηφιακών βοηθών, για να αξιοποιηθεί το μέγιστο πλήθος πληροφορίας. Επιπλέον, οι προεκτάσεις του συστήματος μπορούν να οδηγήσουν σε χρήση του σε εξειδικευμένα συστήματα όπου η πληροφορία που είναι διαθέσιμη είναι περιορισμένη και προέρχεται πάλι από το διαδίκτυο αλλά μόνο από μια συγκεκριμένη πηγή.

Abstract

Artificial Intelligence is a field of technology that is making leaps and bounds every day. It is cutting-edge in solving a multitude of problems and providing high-quality services. Commercial applications in recent years, although with unlimited potential, have focused mainly on Conversational AI or Video and Image Processing AI, resulting in the enthusiasm of the market and users, who are becoming increasingly demanding in terms of the quality and capabilities of their tools.

Chatbots have invaded everyday life in many forms and uses. Their traditional form of answering a closed set of simplistic questions has been replaced by modern systems and applications that suggest, discuss, chat, entertain and advise. Whether as apps on the user's mobile phone or available online, modern assistants are now part of the daily routine and the tools that make it more tolerable and enjoyable. This now universal presence makes it necessary to select these services by adding new elements and parameters, depending on the specificity of each user and their needs.

These applications, being constantly connected to the Internet, have sacrificed user's privacy, making them vulnerable to malware or fraud. At the same time, however, most of these applications cannot make use of a large amount of information available on the Internet even though they are connected to it. Finally, existing systems are not personalised. Their responses and services are rarely tailored to their individual user.

In this thesis an attempt is made to develop a digital question-answering assistant that dynamically searches the web for answers. This system, using pre-trained models and software technologies, answers user's questions by using all the online information. The portability of the system can ensure a private experience as well as answers from specialized data sources.

Experimental results demonstrate that this system is sufficiently accurate while achieving the purposes for which it was developed. The analytical and conversational style of conventional digital assistants is sacrificed in order to exploit the maximum amount of information. Furthermore, the extensions of the system may lead to its use in specialised systems where the information available is limited and again comes from the internet but only from a specific source.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ανδρέα Συμεωνίδη για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα διπλωματική εργασία και για τη στήριξη που μου προσέφερε. Ευχαριστώ επίσης τον υποψήφιο διδάκτορα κ. Νικόλαο Μάλαμα για τη στενή συνεργασία που είχαμε καθ' όλη τη διάρκεια της διπλωματικής και την πολύτιμη βοήθειά του. Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένειά μου που με στήριξε κατά τη διάρκεια αυτής της διπλωματικής αλλά και των σπουδών μου.

Ανάπτυξη Δυναμικού Συστήματος Ερωταπαντήσεων με Πηγή το Διαδίκτυο

Σκαπέτης Χρήστος
skapetis@auth.gr

7 Νοεμβρίου 2023

Περιεχόμενα

1 Εισαγωγή	3
1.1 Περιγραφή του Προβλήματος	4
1.2 Σκοπός - Συνεισφορά της Διπλωματικής Εργασίας	5
1.3 Διάρθρωση Αναφοράς	5
2 Θεωρητικό Υπόβαθρο	6
2.1 Τεχνητή Νοημοσύνη	6
2.2 Chatbots	6
2.3 Γλωσσικά Μοντέλα - Large language models (LLMs)	8
2.4 Transformers	8
2.4.1 Τι είναι ένας transformer	8
2.4.2 Αυτοπροσοχή	9
2.4.3 Multi-head Αυτοπροσοχή	11
2.4.4 Κωδικοποίηση θέσης	11
2.4.5 Κωδικοποιητής	11
2.4.6 Αποκωδικοποιητής	12
2.4.7 Μάσκες	12
2.5 Παρόμοιες Προσεγγίσεις	12
3 Εργαλεία	14
3.1 Python	14
3.1.1 Βιβλιοθήκες	14
3.2 Haystack	15
3.3 Flask	18
4 Μεθοδολογία	19
4.1 Αλγόριθμος	19
4.1.1 Χρήσιμοι Όροι	19
4.1.2 Απάντηση χωρίς αναζήτηση	20
4.1.3 Απάντηση με αναζήτηση με το SERP API	20
4.1.4 Απάντηση με αναζήτηση στο διαδίκτυο	21
4.1.5 Πρώτες παρατηρήσεις	23
4.2 Επιμέρους διαδικασίες	24
4.2.1 Ορίσματα Χρήστη	24
4.2.2 Pipeline	24
4.2.3 Σημαντικές Λέξεις	25
4.2.4 Αξιολόγηση Ιστοσελίδων Βάθους	25
4.3 Μοντέλα	26
4.4 Διεπαφή Χρήστη	28
4.5 Ιδιωτικότητα και Προσωποποιημένο Σύστημα	31
4.6 Προσαρμογή σε ιστοσελίδα	32

5 Πειράματα	33
5.1 Σύγκριση μοντέλων	33
5.1.1 Σύγκριση μοντέλων ως προς την επίδοση	33
5.1.2 Ερωτήματα	33
5.1.3 Ποιότητα των Απαντήσεων	35
5.1.4 Σύγκριση μοντέλων ως προς την επίδοση (άλλες γλώσσες)	42
5.1.5 Σύγκριση μοντέλων ως προς τον χρόνο	43
5.2 Σύγκριση Αξιολόγησης Ιστοσελίδων Βάθους	45
6 Συμπεράσματα	47
6.1 Συμπεράσματα	47
6.2 Μελλοντικές προεκτάσεις	48
Α΄ Ακρωνύμια και συντομογραφίες	49
List of Figures	51

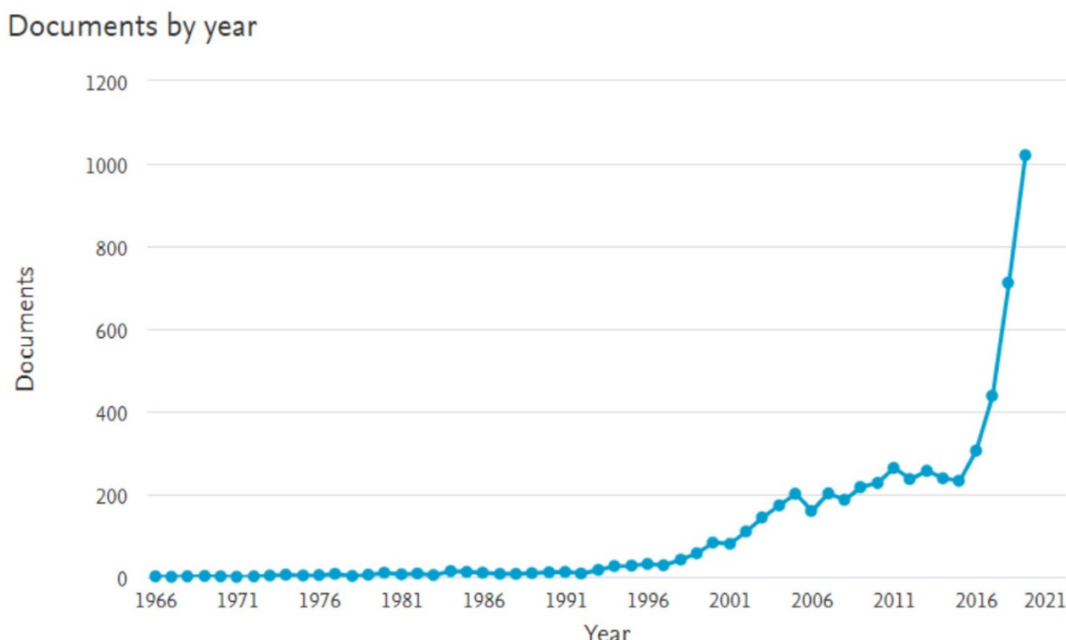
Κεφάλαιο 1

Εισαγωγή

Η Τεχνική Νοημοσύνη (TN) αντιπροσωπεύει μια συναρπαστική πραγματικότητα στον κόσμο της επιστήμης και της τεχνολογίας. Με την εκθετική αύξηση των υπολογιστικών ικανοτήτων και την ανάπτυξη των αλγορίθμων, η TN έχει επιτύχει αξιοσημείωτη πρόοδο σε πολλούς τομείς, ενώ μια από τις πιο σημαντικές εφαρμογές της αφορά την Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP).

Η Επεξεργασία Φυσικής Γλώσσας αναφέρεται στην ικανότητα των υπολογιστών να κατανοούν, να ερμηνεύουν και να παράγουν ανθρώπινη γλώσσα. Αυτό το πεδίο της TN έχει επαναπροσδιορίσει τον τρόπο με τον οποίο αλληλεπιδρούν οι χρήστες με τους υπολογιστές, επιτρέποντας την επικοινωνία μαζί τους με φυσικό τρόπο, χρησιμοποιώντας φωνή ή γραπτό λόγο. Από την αυτόματη μετάφραση κειμένου και την αναγνώριση φωνητικών εντολών έως τη δημιουργία αυτόνομων συστημάτων συνομιλίας, η Επεξεργασία Φυσικής Γλώσσας έχει επαναπροσδιορίσει τη σχέση του ατόμου με τις τεχνολογικές συσκευές και τον ψηφιακό κόσμο γενικότερα.

Το 1950 ο Alan Turing αναρωτήθηκε αν ένα πρόγραμμα υπολογιστή θα μπορούσε να μιλήσει σε μια ομάδα ανθρώπων χωρίς αυτοί να συνειδητοποιήσουν ότι ο συνομιλητής τους είναι τεχνητός [1]. Το 1966 αναπτύχθηκε το πρώτο Chatbot, με το όνομα ELIZA, το οποίο απλώς προσομοίωνε μια συνεδρία ψυχοθεραπευτή επιστρέφοντας τις προτάσεις του χρήστη σε ερωτηματική μορφή [2]. Από τότε η πρόοδος διαφόρων πτυχών της τεχνολογίας έχει αλλάξει ριζικά το τοπίο. Από τη μια, η ριζοσπαστική επανάσταση που εκτόξευσε την υπολογιστική δύναμη που είναι διαθέσιμη, και σε ερευνητικό/βιομηχανικό επίπεδο αλλά και στο μέσο χρήστη, έκανε επιτεύξιμη την υλοποίηση αλγορίθμων και εφαρμογών που απαιτούν τεράστια ποσότητα πόρων [3]. Από την άλλη, η ζήτηση για πιο εξελιγμένα συστήματα που θα παρέχουν υπηρεσίες υψηλού επιπέδου οδήγησε στην ανάπτυξη ψηφιακών βοηθών για πολλές πτυχές της ανθρώπινης καθημερινότητας. Η Siri της Apple, ο IBM Watson της IBM, το Google Assistant της Google, η Cortana της Microsoft και η Alexa της Amazon [4] κατέκλυσαν εκατομμύρια νοικοκυριά σε όλον τον κόσμο στα τέλη της δεκαετίας του 2010 [5] και από πολυτέλεια ή ένα gadget πολυτελείας έγιναν αναγκαιότητα και εργαλεία της καθημερινότητας. Ωστόσο, οι απαιτήσεις για ακόμα καλύτερα μοντέλα και υπηρεσίες δημιούργησαν την ανάγκη για αποτελεσματικότερα και πιο εξειδικευμένα μοντέλα και εφαρμογές. Ενδεικτική της ζήτησης στο συγκεκριμένο τομέα είναι και η κατακόρυφη αύξηση ερευνητικών άρθρων και δημοσιεύσεων με αντικείμενο τα Chatbots και γενικότερα τους πράκτορες συνομιλίας, όπως φαίνεται και στο Σχήμα 1.1.



Σχήμα 1.1: Αποτελέσματα αναζήτησης στο Scopus Από το 1966 μέχρι το 2021 με λέξεις κλειδιά “chatbot” ή “conversation agent” ή “conversational interface” [4]

Προφανής είναι και η εξέλιξη σε πιο τεχνικό επίπεδο. Στο παρελθόν χρησιμοποιούνταν αναπαραστάσεις ενός επιπέδου με διανύσματα λέξεων που εισάγονταν σε task-specific αρχιτεκτονικές, ενώ αργότερα επιλέχθηκαν RNNs με πολλαπλά επίπεδα από αναπαραστάσεις και νοηματικές καταστάσεις. Πλέον έχουν κυριαρχήσει προ-εκπαιδευμένες γλωσσικές αναπαραστάσεις σε NLP συστήματα που εφαρμόζονται σε αυξανόμενα ελαστικά task-agnostic συστήματα. Προεκπαιδευμένα μοντέλα ή μοντέλα γλωσσικών μετασχηματισμών ρυθμίζονται λεπτομερώς (fine-tuned) αφαιρώντας την ανάγκη για αρχιτεκτονικές με σαφείς στόχους [6].

1.1 Περιγραφή του Προβλήματος

Πρόσφατα πραγματοποιήθηκε πληθώρα προσπαθειών να αναπτυχθούν Chatbots τα οποία μερικές φορές ακόμα και υπερβαίνουν τις ανθρώπινες δυνατότητες σε ορισμένους γλωσσικούς τομείς. Τα περισσότερα από αυτά βασίστηκαν σε LLMs, δηλαδή μεγάλα γλωσσικά μοντέλα, με κυριότερα την 3η και 4η γενιά του μοντέλου GPT δύο μεγάλων πολυτροπικών μοντέλων γλώσσας που δημιουργήθηκαν από την εταιρία OpenAI [6] [7]. Κύριο αρνητικό χαρακτηριστικό τους είναι η μη ανανέωση της πληροφορίας που έχουν αποθηκεύσει, η στατικότητα αυτής, αλλά και η αδυναμία σύνδεσής τους στο Διαδίκτυο για εύρεση λιγότερο γνωστών απαντήσεων (πχ. ChatGPT).

Επίσης, το σύνολο της πληροφορίας που χρησιμοποιείται δεν έχει προσωποποιημένο χαρακτήρα, με όλους τους χρήστες να μοιράζονται τα ίδια κειμενικά δεδομένα. Από την άλλη, η Google ανέπτυξε το Bard και η Microsoft το Bing που επιλύουν την αδυναμία της offline λειτουργίας, θυσιάζοντας ποιότητα στις απαντήσεις τους. Ταυτόχρονα, αυτά τα μοντέλα, λόγω της συνεχούς σύνδεσής τους στο διαδίκτυο, θυσιάζουν την ιδιωτικότητα του χρήστη. Όλες οι αναζητήσεις του χρήστη, καθώς και τα αποτελέσματα που του επιστρέφουν οι διάφορες εφαρμογές, αφήνουν ένα ψηφιακό αποτύπωμα που μερικές φορές όχι μόνο δεν είναι επιθυμητό, αλλά είναι και πηγή προβλημάτων [8].

Παράλληλα, υπάρχουν και εξειδικευμένα Chatbots που εστιάζουν σε μια μόνο υπηρεσία. Λειτουργούν είτε σαν συνομιλίες πραγματικού χρόνου (live chats) σε τομείς όπως η εξυπηρέτηση πελατών, είτε σαν υπηρεσίες συχνών ερωτήσεων σε οργανισμούς [9]. Αυτές οι εφαρμογές βασίζονται σε τεχνικές αντιστοίχισης προτύπων, σε συστήματα κανόνων και σε προαποφασισμένες απαντήσεις [4]. Όπως είναι προφανές, αυτές οι τεχνικές

συνήθως παρέχουν μειωμένης ποιότητας απαντήσεις και στοχεύουν στην εξυπηρέτηση των χρηστών μόνο στα πιο βασικά ερωτήματα και προβλήματα.

1.2 Σκοπός - Συνεισφορά της Διπλωματικής Εργασίας

Στην παρούσα διπλωματική εργασία σχεδιάζεται και υλοποιείται ένα ιδιωτικό σύστημα ψηφιακού βοηθού ερωταπαντήσεων (Question Answering) με ιδιαίτερα προσωποποιημένο χαρακτήρα. Για την εύρεση της απάντησης στο ερώτημα του χρήστη γίνεται χρήση πληροφορίας που αναζητήθηκε δυναμικά στο διαδίκτυο. Στην συνέχεια, με την αξιοποίηση τεχνικών και προεκπαιδευμένων μοντέλων μηχανικής μάθησης, πραγματοποιείται εξαγωγή της απάντησης από το σύνολο της πληροφορίας αυτής. Το σύστημα επιστρέφει στον χρήστη μια απάντηση σε μορφή κειμένου, ένα σκορ συνάφειας (αξιολόγηση) της απάντησης, την πηγή-ιστοσελίδα που χρησιμοποιήθηκε, καθώς και το πλαίσιο (context) στο οποίο βρέθηκε η απάντηση αυτή.

Τα μοντέλα που χρησιμοποιούνται για την εξαγωγή της απάντησης είναι 3 προεκπαιδευμένα μοντέλα ερωταπαντήσεων στην Αγγλική γλώσσα, RoBERTa [10], MiniLM [11] και ALBERT (XXL) [12], και 3 πολυγλωσσικά (multilingual) προεκπαιδευμένα μοντέλα το XLM RoBERTa [13], το DeBERTa V3 [14] και το Gelectra [15]. Παράλληλα γίνεται χρήση δύο διαφορετικών Readers, του FARMReader [16] και του TransformersReader [16].

Ο απώτερος σκοπός είναι η δημιουργία ενός συστήματος ιδιωτικού και προσωποποιημένου που θα διατηρεί ως έναν βαθμό την ποιότητα των προϋπαρχόντων συστημάτων. Τα δεδομένα που χρησιμοποιούνται για την εξαγωγή των απαντήσεων δεν είναι πλέον στατικά, αλλά ανανεώνονται δυναμικά κάθε φορά που ο χρήστης υποβάλει ένα νέο ερώτημα στο σύστημα. Σε αντίθεση με τα προϋπάρχοντα συστήματα, το σύνολο της πληροφορίας ανανεώνεται συνεχώς κατά τη χρήση της εφαρμογής. Επιπρόσθετα, η δυνατότητα εξαγωγής της πηγής γνώσης με σκοπό την λειτουργία του συστήματος σε εκτός σύνδεσης (offline) λειτουργία, καθιστά το εν λόγω εγχείρημα μια προσπάθεια προστασίας της ιδιωτικότητας, τόσο των ερωτήσεων, όσο και των πηγών και των απαντήσεων.

1.3 Διάρθρωση Αναφοράς

Η διάρθρωση της παρούσας διπλωματικής εργασίας είναι η εξής:

- Κεφάλαιο 1: Εισαγωγή στη διπλωματική εργασία και παρουσίαση του αντικειμενικού σκοπού της και πρόσφατων προσεγγίσεων στον χώρο.
- Κεφάλαιο 2: Παρουσίαση του θεωρητικού υπόβαθρου της εργασίας, κυρίως γύρω από τις έννοιες της μηχανικής μάθησης και των μετασχηματισμών.
- Κεφάλαιο 3: Παρουσίαση των διαφόρων εργαλείων λογισμικού που χρησιμοποιήθηκαν κατά την εκπόνηση της διπλωματικής εργασίας.
- Κεφάλαιο 4: Παρουσίαση της μεθοδολογίας της εφαρμογής που αναπτύχθηκε καθώς και του περιβάλλοντος αυτής.
- Κεφάλαιο 5: Παρουσίαση διαφόρων συγκρίσεων που πραγματοποιήθηκαν κατά τη μελέτη αυτής της διπλωματικής εργασίας.
- Κεφάλαιο 6: Παρουσίαση συμπερασμάτων αυτής της διπλωματικής εργασίας και μελλοντικών πιθανών προεκτάσεων πάνω σε αυτή.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Σε αυτό το κεφάλαιο γίνεται μια σύντομη αναφορά στο θεωρητικό υπόβαθρο και στις έννοιες που χρειάζονται για την κατανόηση της παρούσας διπλωματικής εργασίας. Αρχικά, παρουσιάζονται γενικότερες έννοιες και στην συνέχεια αναλύονται συγκεκριμένοι αλγόριθμοι και μεθοδολογίες πάνω στα οποία βασίστηκε η εργασία.

2.1 Τεχνητή Νοημοσύνη

Τεχνητή νοημοσύνη (TN) - Artificial Intelligence (AI) είναι η νοημοσύνη -αντίληψη, σύνθεση και εξαγωγή συμπερασμάτων- που επιδεικνύεται από μηχανές, σε αντίθεση με τη νοημοσύνη που επιδεικνύουν οι άνθρωποι ή άλλα ζώα. Παραδείγματα εργασιών στις οποίες γίνεται αυτό περιλαμβάνουν την αναγνώριση ομιλίας, την όραση υπολογιστών, τη μετάφραση μεταξύ (φυσικών) γλωσσών και άλλα [17].

Οι εφαρμογές τεχνητής νοημοσύνης περιλαμβάνουν προηγμένες μηχανές αναζήτησης στο διαδίκτυο, συστήματα συστάσεων, κατανόηση της ανθρώπινης ομιλίας, αυτόνομη οδήγηση, γλωσσικά μοντέλα και ψηφιακούς βοηθούς (το θέμα της παρούσης εργασίας), αυτοματοποιημένη λήψη αποφάσεων, και ανταγωνισμό στο υψηλότερο επίπεδο σε συστήματα στρατηγικών παιχνιδιών.

Καθώς οι μηχανές γίνονται όλο και πιο ικανές, οι εργασίες που θεωρούνται ότι απαιτούν “νοημοσύνη” συχνά αφαιρούνται από τον ορισμό της τεχνητής νοημοσύνης, ένα φαινόμενο που είναι γνωστό ως το φαινόμενο της τεχνητής νοημοσύνης [18].

Τα διάφορα επιμέρους πεδία της έρευνας της TN εστιάζουν σε συγκεκριμένους στόχους με τη χρήση ορισμένων τεχνικών και εργαλείων. Οι παραδοσιακοί στόχοι της έρευνας της TN περιλαμβάνουν τη λογική, την αναπαράσταση γνώσης, τη μάθηση, την επεξεργασία φυσικής γλώσσας, την αντίληψη και την ικανότητα κίνησης και χειρισμού αντικειμένων [19]. Για την επίλυση αυτών των προβλημάτων, οι ερευνητές της TN έχουν προσαρμόσει και ενσωματώσει ένα ευρύ φάσμα τεχνικών επίλυσης προβλημάτων, συμπεριλαμβανομένων της αναζήτησης και της μαθηματικής βελτιστοποίησης, της τυπικής λογικής, των νευρωνικών δικτύων και των μεθόδων που βασίζονται στη στατιστική, τις πιθανότητες και τα οικονομικά. Ταυτόχρονα, η TN χρησιμοποιεί τεχνικές από την επιστήμη των υπολογιστών, την ψυχολογία, τη γλωσσολογία, τη φιλοσοφία και πολλούς άλλους τομείς.

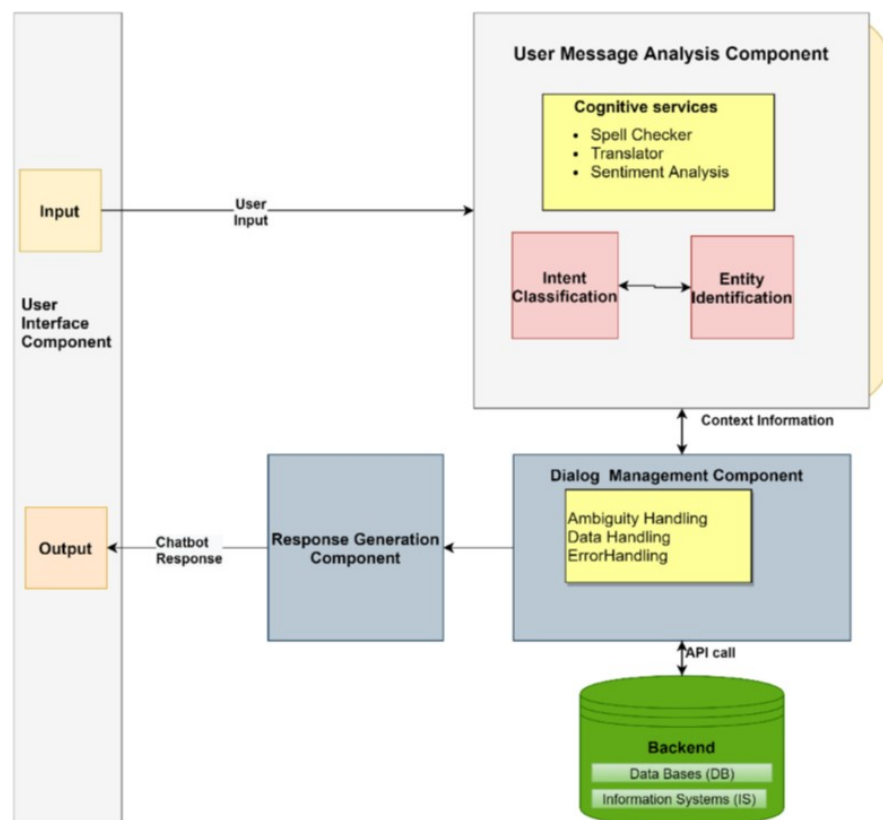
2.2 Chatbots

Ένα chatbot, στο πιο βασικό επίπεδο, είναι ένα πρόγραμμα υπολογιστή που προσομοιώνει και επεξεργάζεται την ανθρώπινη συνομιλία (είτε γραπτή είτε προφορική), επιτρέποντας στους ανθρώπους να αλληλεπιδρούν με ψηφιακές συσκευές σαν να επικοινωνούσαν με ένα πραγματικό πρόσωπο. Τα chatbots μπορεί να είναι τόσο απλά όσο και υποτυπώδη προγράμματα που απαντούν σε ένα απλό ερώτημα με μια μονολεκτική απάντηση,

ή τόσο εξελιγμένα όσο και ψηφιακοί βοηθοί που μαθαίνουν και εξελίσσονται, ώστε να παρέχουν αυξανόμενα επίπεδα εξατομίκευσης καθώς συλλέγουν και επεξεργάζονται πληροφορίες [9].

Με τη χρήση της τεχνητής νοημοσύνης, αυτοματοποιημένων κανόνων, της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) και της μηχανικής μάθησης (ML), τα chatbots επεξεργάζονται δεδομένα, για να παρέχουν απαντήσεις σε κάθε είδους αιτήματα. Υπάρχουν δύο κύρια είδη chatbots. Τα δηλωτικά και τα συνομιλητικά. Τα δηλωτικά chatbots είναι προγράμματα με έναν συγκεκριμένο σκοπό. Επικεντρώνονται στην εκτέλεση μιας συγκεκριμένης λειτουργίας, χρησιμοποιώντας κυρίως κανόνες και NLP και πολύ λίγη μηχανική μάθηση. Παράγουν αυτοματοποιημένες απαντήσεις ως απόκριση στη συνομιλία με τον χρήστη. Οι αλληλεπιδράσεις αυτές είναι πολύ συγκεκριμένες και δομημένες, με κύρια εφαρμογή την εξυπηρέτηση και υποστήριξη πελατών. Από την άλλη, τα συνομιλητικά αναφέρονται συχνά και ως ψηφιακοί βοηθοί και είναι πιο εξελιγμένα και διαδραστικά. Αυτά τα chatbots έχουν επίγνωση του πλαισίου λειτουργίας τους και χρησιμοποιούν την κατανόηση φυσικής γλώσσας (Natural-language understanding - NLU), τα νευρωνικά δίκτυα και τη μηχανική μάθηση για συνεχή εκπαίδευσή τους. Οι ψηφιακοί βοηθοί μπορούν να μαθαίνουν τις προτιμήσεις ενός χρήστη με την πάροδο του χρόνου, να παρέχουν συστάσεις, ακόμη και να προβλέπουν ανάγκες. Εκτός από την παρακολούθηση των δεδομένων και των προθέσεων, μπορούν να ξεκινούν συζητήσεις. Η Siri της Apple και η Alexa της Amazon είναι παραδείγματα chatbots με προσανατολισμό στον καταναλωτή, βασισμένα σε δεδομένα και προβλέψεις [9].

Η πιο συνήθης αρχιτεκτονική για chatbots που λειτουργούν με κανόνες αποτελείται από συγκεκριμένα στοιχεία. Τη διεπαφή του χρήστη, τον χειριστή ανάλυσης μηνυμάτων, έναν χειριστή διαλόγου, το backend και ένα στοιχείο παραγωγής απαντήσεων. Στο Σχήμα 2.1 παρουσιάζεται η βασική δομή αυτής της τυπικής αρχιτεκτονικής.



Σχήμα 2.1: Τυπική Αρχιτεκτονική ενός chatbot [4]

2.3 Γλωσσικά Μοντέλα - Large language models (LLMs)

Ένα μεγάλο γλωσσικό μοντέλο (LLM) είναι ένα γλωσσικό μοντέλο που αποτελείται από ένα νευρωνικό δίκτυο με μεγάλο αριθμό παραμέτρων (γενικά της τάξης του ενός δισεκατομμυρίου ή και περισσότερων βαρών) [20]. Η εξάσκηση του μοντέλου έχει γίνει με μεγάλες ποσότητες μη-σημασμένου κειμένου με τη χρήση αυτοεποπτευόμενης ή ημιεποπτευόμενης μάθησης. Τα LLMs εμφανίστηκαν γύρω στο 2018 και αποδίδουν καλά σε μια ευρεία ποικιλία εργασιών. Αυτό έχει μετατοπίσει το επίκεντρο της έρευνας για την επεξεργασία της φυσικής γλώσσας από το προηγούμενο πρότυπο της εκπαίδευσης εξειδικευμένων εποπτευόμενων μοντέλων για συγκεκριμένες εργασίες [21].

Η κυρίαρχη προσέγγιση για την περιγραφή του νόηματος, όχι μόνο στη γλωσσολογία και στην φιλοσοφία της γλώσσας, αλλά και για τις γλώσσες προγραμματισμού, είναι η προσέγγιση της σημασιολογίας ή της θεωρίας της αναφοράς: το νόημα μιας λέξης, φράσης ή πρότασης είναι το σύνολο των αντικειμένων ή καταστάσεων του κόσμου που περιγράφει (ή μια μαθηματική αφαίρεσή τους). Αυτό έρχεται σε αντίθεση με την απλή διανεμημένη σημασιολογία (ή θεωρία της χρήσης του νόηματος) της σύγχρονης εμπειρικής εργασίας στο NLP, σύμφωνα με την οποία το νόημα μιας λέξης είναι απλώς μια περιγραφή των συμφοραζομένων στα οποία εμφανίζεται. Ορισμένοι έχουν προτείνει ότι η τελευταία δεν είναι καθόλου θεωρία της σημασιολογίας αλλά απλώς μια αναμάσηση διανεμημένων ή συντακτικών γεγονότων [21].

2.4 Transformers

2.4.1 Τι είναι ένας transformer

Ο μετασχηματιστής (transformer) είναι ένα νέο είδος νευρωνικής αρχιτεκτονικής που κωδικοποιεί τα δεδομένα εισόδου ως ισχυρά χαρακτηριστικά μέσω του μηχανισμού προσοχής [22].

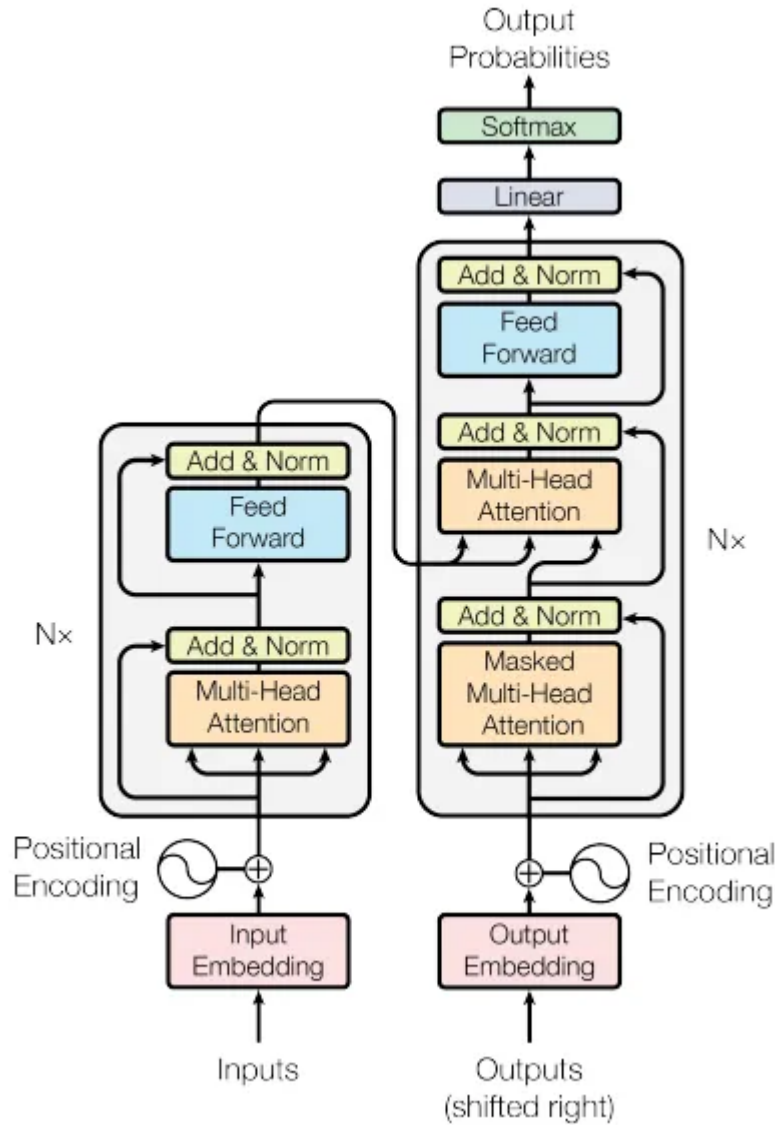
Διακρίνεται από την υιοθέτηση της αυτοπροσοχής, σταθμίζοντας διαφορετικά τη σημασία κάθε μέρους των δεδομένων εισόδου (που περιλαμβάνει την αναδρομική έξοδο). Χρησιμοποιείται κυρίως στους τομείς της επεξεργασίας φυσικής γλώσσας (NLP) και της όρασης υπολογιστών (CV) [23].

Όπως και τα αναδρομικά νευρωνικά δίκτυα (RNN), οι μετασχηματιστές έχουν σχεδιαστεί για την επεξεργασία διαδοχικών δεδομένων εισόδου, όπως η φυσική γλώσσα, με εφαρμογές σε εργασίες όπως η μετάφραση και η περίληψη κειμένου. Ωστόσο, σε αντίθεση με τα RNN, οι μετασχηματιστές επεξεργάζονται ολόκληρη την είσοδο με τη μία. Ο μηχανισμός προσοχής παρέχει πλαίσιο για οποιαδήποτε θέση στην ακολουθία εισόδου. Για παράδειγμα, εάν τα δεδομένα εισόδου είναι μια πρόταση φυσικής γλώσσας, ο μετασχηματιστής δεν χρειάζεται να επεξεργάζεται μία λέξη κάθε φορά. Αυτό επιτρέπει μεγαλύτερο παραλληλισμό από τα RNN και, επομένως, μειώνει τους χρόνους εκπαίδευσης [24].

Σε προβλήματα sequence-to-sequence, όπως η νευρωνική μετάφραση, οι αρχικές προτάσεις βασίστηκαν στη χρήση RNN σε μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή. Αυτές οι αρχιτεκτονικές έχουν έναν μεγάλο περιορισμό, όταν εργάζονται με μεγάλες ακολουθίες. Η ικανότητά τους να διατηρούν πληροφορίες από τα πρώτα στοιχεία χανόταν, όταν νέα στοιχεία ενσωματώνονταν στην ακολουθία. Στον κωδικοποιητή, η κρυφή κατάσταση σε κάθε βήμα συνδέεται με μια συγκεκριμένη λέξη στην πρόταση εισόδου, συνήθως μια από τις πιο πρόσφατες. Επομένως, αν ο αποκωδικοποιητής έχει πρόσβαση μόνο στην τελευταία κρυφή κατάσταση του κωδικοποιητή, θα χάσει σχετικές πληροφορίες για τα πρώτα στοιχεία της ακολουθίας. Στη συνέχεια, για να αντιμετωπιστεί αυτός ο περιορισμός, εισήχθη μια νέα έννοια, ο μηχανισμός προσοχής.

Στους μετασχηματιστές (Transformers) (Σχήμα 2.2), αντί να δίνεται προσοχή στην τελευταία κατάσταση του κωδικοποιητή, όπως γίνεται συνήθως με τα RNN, σε κάθε βήμα του αποκωδικοποιητή εξετάζονται όλες οι καταστάσεις του κωδικοποιητή, έχοντας πρόσβαση σε πληροφορίες για όλα τα στοιχεία της ακολουθίας εισόδου. Αυτό κάνει η προσοχή, εξάγει πληροφορίες από ολόκληρη την ακολουθία, ένα σταθμισμένο άθροισμα όλων των προηγούμενων καταστάσεων του κωδικοποιητή. Αυτό επιτρέπει στον αποκωδικοποιητή να αποδίδει μεγαλύτερο βάρος ή σημασία σε ένα συγκεκριμένο στοιχείο της εισόδου για κάθε στοιχείο της εξόδου. Μαθαίνει σε κάθε βήμα να εστιάζει στο σωστό στοιχείο της εισόδου, για να προβλέψει το επόμενο στοιχείο της εξόδου. Αλλά αυτή η προσέγγιση συνεχίζει να έχει έναν σημαντικό περιορισμό, κάθε ακολουθία πρέπει να αντιμετωπίζεται ένα στοιχείο κάθε φορά. Τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής πρέπει να περιμένουν

μέχρι την ολοκλήρωση των $t - 1$ βημάτων, για να επεξεργαστούν το t -ωστό βήμα. Έτσι, όταν πρόκειται για τεράστια σώματα κειμένου, η διαδικασία είναι πολύ χρονοβόρα και υπολογιστικά αναποτελεσματική.



Σχήμα 2.2: Το μοντέλο αρχιτεκτονικής Transformer [24]

2.4.2 Αυτοπροσοχή

Η Αυτοπροσοχή (self-attention) είναι μια sequence-to-sequence λειτουργία: η είσοδος είναι μια ακολουθία διανυσμάτων και η έξοδος είναι μια ακολουθία διανυσμάτων. Έστω τα διανύσματα εισόδου x_1, x_2, \dots, x_t και τα αντίστοιχα διανύσματα εξόδου y_1, y_2, \dots, y_t . Όλα τα διανύσματα έχουν διάσταση k . Για να παραχθεί το διάνυσμα εξόδου y_i , η λειτουργία της αυτοπροσοχής παίρνει απλά έναν σταθμισμένο μέσο όρο σε όλα τα διανύσματα εισόδου, η απλούστερη επιλογή είναι το εσωτερικό γινόμενο.

Ερώτημα, κλειδί και τιμή

Κάθε διάνυσμα εισόδου χρησιμοποιείται με τρεις διαφορετικούς τρόπους στον μηχανισμό αυτοπροσοχής: το ερώτημα, το κλειδί και η τιμή. Σε κάθε ρόλο, συγκρίνεται με τα άλλα διανύσματα για να πάρει τη δική του

έξοδο y_i (ερώτημα), για να πάρει την j -οστή έξοδο y_j (κλειδί) και για να υπολογίσει κάθε διάνυσμα εξόδου, αφού καθοριστούν τα βάρη (τιμή).

Για να αποκτήσουμε αυτούς τους ρόλους, χρειαζόμαστε τρεις πίνακες βαρών (Σχήμα 2.3) διαστάσεων $k \times k$ και υπολογίζουμε τρεις γραμμικούς μετασχηματισμούς για κάθε x_i :

$$q_i = W_q x_i \quad k_i = W_k x_i \quad v_i = W_v x_i$$

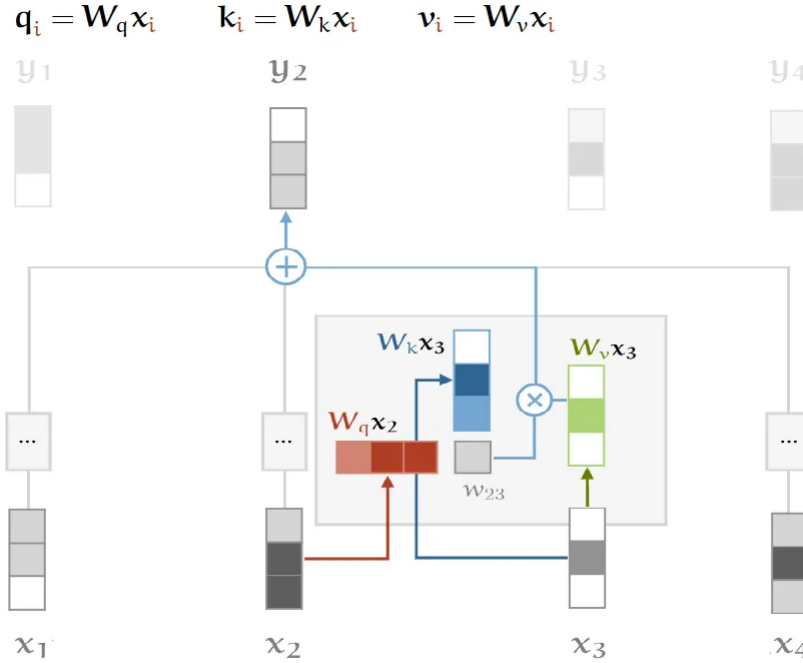


Illustration of the self-attention with **key**, **query** and **value**

Σχήμα 2.3: Αυτοπροσοχή

Αυτοί οι τρεις πίνακες είναι συνήθως γνωστοί ως K , Q και V , τρία μαθησιακά επίπεδα βαρών που εφαρμόζονται στην ίδια κωδικοποιημένη είσοδο.

Γραμμωτό (Scaled) Εσωτερικό Γινόμενο

Στη συνέχεια, γίνεται η χρήση των τριών πινάκων για την εύρεση των σκορ προσοχής. Το σκορ προσοχής είναι μια βαθμολογία που μετράει πόση έμφαση πρέπει να δοθεί σε άλλες θέσεις ή λέξεις της ακολουθίας εισόδου σε σχέση με μια λέξη σε μια συγκεκριμένη θέση. Δηλαδή, το εσωτερικό γινόμενο του διανύσματος του ερωτήματος με το διάνυσμα-κλειδί της αντίστοιχης λέξης που βαθμολογείται. Για παράδειγμα, στη θέση 1 υπολογίζεται το εσωτερικό γινόμενο των q_1 και k_1 μετά των q_1 και k_2 , q_1 και k_3 κ.ο.κ.

Για αυτήν τη διαδικασία εφαρμόζεται ένα συντελεστής κλιμάκωσης, στη συγκεκριμένη περίπτωση ο softmax:

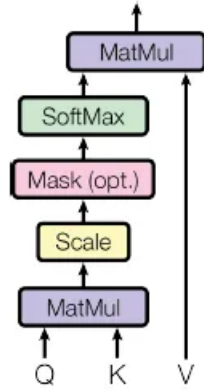
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad [24]$$

Η συνάρτηση softmax δεν μπορεί να λειτουργήσει σωστά με μεγάλες τιμές, με αποτέλεσμα να εξαφανίζονται οι κλίσεις και να επιβραδύνεται η μάθηση. Μετά το "softmaxing" πολλαπλασιάζεται με τον πίνακα των τιμών, για να παραμείνουν οι τιμές των λέξεων στις οποίες είναι επιθυμητό να εστιάσει ο αλγόριθμος, και ελαχιστοποιούνται ή αφαιρούνται οι τιμές για τις άσχετες λέξεις (η τιμή του στον πίνακα V πρέπει να είναι πολύ μικρή).

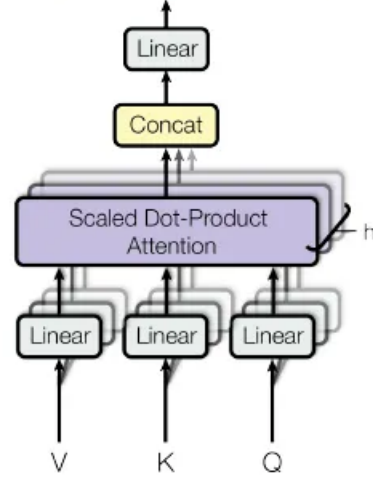
2.4.3 Multi-head Αυτοπροσοχή

Μπορούμε να δώσουμε στην αυτοπροσοχή μεγαλύτερη δύναμη διάκρισης, συνδυάζοντας πολλαπλά heads αυτοπροσοχής (Σχήμα 2.4), διαιρώντας τα διανύσματα λέξεων με έναν σταθερό αριθμό (h , αριθμός heads) τεμαχίων, και στη συνέχεια η αυτοπροσοχή εφαρμόζεται στα αντίστοιχα τεμάχια, χρησιμοποιώντας τους υποπίνακες Q , K και V [24].

Scaled Dot-Product Attention



Multi-Head Attention



Σχήμα 2.4: (αριστερά) Γραμμωτό Εσωτερικό Γινόμενο. (δεξιά) Multi-head Αυτοπροσοχή αποτελούμενη από παράλληλα layers [24]

Ακολουθώς, συνενώνονται οι πίνακες σε έναν διευρυμένο πίνακα και πολλαπλασιάζονται με τα βάρη w_0 . Ο τελικός πίνακας περιέχει πληροφορία από όλα τα heads προσοχής [24].

2.4.4 Κωδικοποίηση θέσης

Με την μέχρι τώρα μέθοδο, οι ίδιες λέξεις σε διαφορετική σειρά δίνουν ακριβώς το ίδιο αποτέλεσμα, οπότε τίθεται η ανάγκη αναπαράστασης της θέσης της λέξης στην πρόταση. Επιλέγεται μια συνάρτηση για την απεικόνιση της θέσης αυτής και προστίθεται σε ένα διάνυσμα. Στην εργασία των Vaswani, et al., 2017 χρησιμοποιείται η ημιτονοειδής συνάρτηση [24]:

$$PE_{pos, 2i} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{pos, 2i+1} = \cos(pos/10000^{2i/d_{model}})$$

2.4.5 Κωδικοποιητής

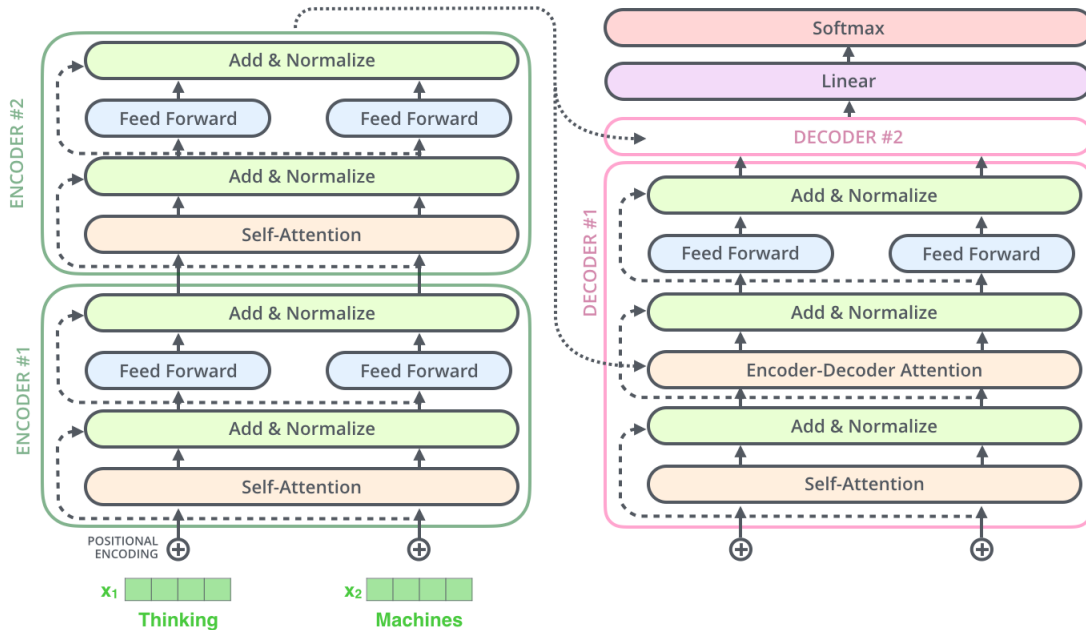
Ο κωδικοποιητής έχει 4 στοιχεία. Την κωδικοποίηση θέσης, 8 επίπεδα RNN, μια σύνδεση υπολοίπου και ένα επίπεδο κανονικοποίησης. Το νευρωνικό δίκτυο αποτελείται από 6 πανομοιότυπα επίπεδα και 2 υποεπίπεδα. Έναν μηχανισμό αυτοπροσοχής πολλαπλών heads και ένα δίκτυο τροφοδότησης:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Στη συνέχεια, υπάρχει ένα υπόλοιπο που συνδέεται σε κάθε υποεπίπεδο, και στο τέλος εφαρμόζεται κανονικοποίηση.

2.4.6 Αποκωδικοποιητής

Αντίστοιχα με τον κωδικοποιητή (Σχήμα 2.5 Αριστερά), ο αποκωδικοποιητής (Σχήμα 2.5 Δεξιά) έχει τα ίδια στοιχεία. Την κωδικοποίηση θέσης, 9 επίπεδα RNN, την αφαίρεση του υπολοίπου και την κανονικοποίηση.



Σχήμα 2.5: Κωδικοποιητής και Αποκωδικοποιητής [25]

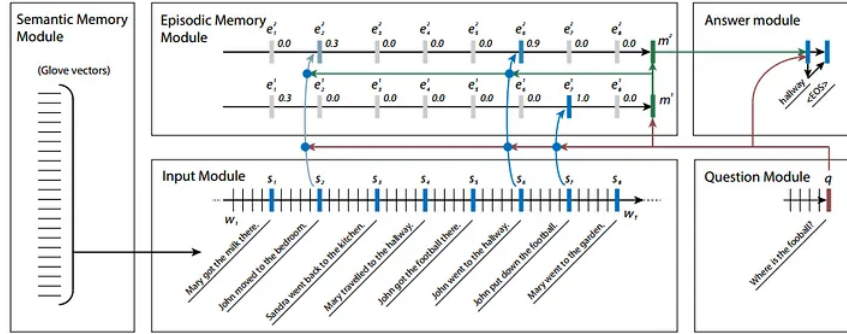
Στο τέλος των N στοιβαγμένων αποκωδικοποιητών, το γραμμικό επίπεδο, ένα πλήρως συνδεδεμένο δίκτυο, μετατρέπει τις στοιβαγμένες εξόδους σε ένα πολύ μεγαλύτερο διάνυσμα, τα logits. Το στρώμα softmax μετατρέπει στη συνέχεια αυτές τις βαθμολογίες (logits) σε πιθανότητες (όλες θετικές, όλες αθροίζουν 1,0). Το κελί με την υψηλότερη πιθανότητα επιλέγεται και η λέξη που σχετίζεται με αυτό παράγεται ως εξόδος για αυτό το χρονικό βήμα [25].

2.4.7 Μάσκες

Τα τρία αυτά μέρη, softmax, κωδικοποιητής και αποκωδικοποιητής, ενώνονται με τη βοήθεια τριών масκών. Η πρώτη μάσκα είναι η μάσκα κωδικοποίησης η οποία αφαιρεί τα pad tokens από τον υπολογισμό της προσοχής. Αν η επεξεργασία των δεδομένων γίνεται σε πακέτα N στοιχείων και η ακολουθία μας έχει M στοιχεία (με $M < N$), τότε τα pad tokens προστίθενται, ώστε να καλυφθεί το ελάχιστο απαιτούμενο μήκος. Στη συνέχεια, η πρώτη μάσκα του αποκωδικοποιητή ενώνει τις μάσκες συμπλήρωσης και look ahead, και θα βοηθήσει στην απόρριψη των tokens στο μέλλον, ενώ η δεύτερη είναι η μάσκα συμπλήρωσης και εφαρμόζεται στο επίπεδο προσοχής κωδικοποιητή-αποκωδικοποιητή.

2.5 Παρόμοιες Προσεγγίσεις

Αυτήν την στιγμή είναι διαθέσιμη πληθώρα ψηφιακών βοηθών. Η βασική μεθοδολογία που χρησιμοποιείται για την ανάπτυξη τέτοιων συστημάτων είναι τα δίκτυα δυναμικής μνήμης, Dynamic Memory Network (DMN) [26]. Ο λόγος για αυτήν την επιλογή είναι δυϊκός. Επιτυγχάνεται η βέλτιστη ταχύτητα για διαδικασία ερωταπαντήσεων και NLP, ενώ την ίδια στιγμή η τμηματική αρχιτεκτονική (βλέπε Σχήμα 2.6) των DMN επιτρέπει την εφαρμογή διαφορετικών συστημάτων.



Σχήμα 2.6: Αρχιτεκτονική δικτύων δυναμικής μνήμης [26]

Η πλειοψηφία των πιο δημοφιλών μοντέλων και εφαρμογών είναι προεκπαιδευμένα μεγάλα γλωσσικά μοντέλα τα οποία δεν μπορούν να αναζητήσουν πληροφορία στο διαδίκτυο. Εφαρμογές όπως το ChatGPT έχουν εξαιρετικές επιδόσεις, αλλά αδυνατούν να απαντήσουν ερωτήματα σε πραγματικό χρόνο. Όλες οι γνώσεις που έχουν αποκτήσει, προέκυψαν από την εκπαίδευσή τους στο παρελθόν.

Ωστόσο, η Microsoft με το Bing Chat και η Google με το Google Bard κάλυψαν αυτό το κενό και έχουν άμεση πρόσβαση σε ολόκληρο το διαδίκτυο. Αυτό το οποίο αδυνατούν όμως να επιλύσουν είναι η εξειδίκευση και η ιδιωτικότητα. Αρχικά, δεν μπορεί ο χρήστης να ζητήσει να χρησιμοποιηθούν ορισμένες πηγές για την απάντηση των ερωτημάτων του. Σε δεύτερο πλαίσιο δεν μπορούν τα συστήματα αυτά να χρησιμοποιηθούν για ερωταπαντήσεις πάνω σε συγκεκριμένες βάσεις δεδομένων, ιστοσελίδες ή άλλο περιεχόμενο πραγματικού χρόνου. Όσον αφορά την ιδιωτικότητα, η αναζήτηση στο διαδίκτυο αφήνει συνεχώς αποτυπώματα και κάθε prompt αποθηκεύεται από την εκάστοτε εταιρία.

Μια άλλη προσέγγιση είναι το hubspot [27]. Αυτή η πλατφόρμα είναι εξειδικευμένη στους ψηφιακούς βοηθούς για επιχειρήσεις. Επιτρέπει στους χρήστες να χτίζουν εξειδικευμένους ψηφιακούς βοηθούς πάνω σε μια θεματολογία, κυρίως πάνω στα δεδομένα μιας επιχείρησης. Αυτή η υλοποίηση δεν έχει πρόσβαση σε πληροφορία στον ιστό, αλλά καλύπτει το δεύτερο τμήμα αυτής της διπλωματικής που είναι η προσωποποιημένη εφαρμογή.

Κεφάλαιο 3

Εργαλεία

3.1 Python

Η Python είναι διερμηνευόμενη (interpreted), γενικού σκοπού (general-purpose) και υψηλού επιπέδου, γλώσσα προγραμματισμού. Η φιλοσοφία σχεδιασμού της δίνει έμφαση στην αναγνωσιμότητα του κώδικα με τη χρήση σημαντικής εσοχής μέσω του κανόνα off-side. Ανήκει στις γλώσσες προστακτικού προγραμματισμού (Imperative programming) και υποστηρίζει τόσο το διαδικαστικό (procedural programming) όσο και το αντικειμενοστρεφές (object-oriented programming) προγραμματιστικό υπόδειγμα (programming paradigm). Είναι δυναμική γλώσσα προγραμματισμού (dynamically typed) και υποστηρίζει συλλογή απορριμμάτων (garbage collection) [28]. Σήμερα, χρησιμοποιείται ιδιαίτερα στη μηχανική μάθηση, λόγω της ύπαρξης μιας μεγάλης και ενεργής κοινότητας, αλλά και της ανάπτυξης πολλών σχετικών με τον τομέα βιβλιοθηκών, όπως η TensorFlow, η PyTorch, η Scikit-learn και άλλες.

3.1.1 Βιβλιοθήκες

Κατά τη διάρκεια της υλοποίησης αυτής της διπλωματικής εργασίας έγινε χρήση πληθώρας βιβλιοθηκών στην python. Οι πιο γνωστές και απλές, όπως οι numpy, os, pandas, tabulate, παραλείπονται. Ταυτόχρονα, βιβλιοθήκες όπως TensorFlow [29] και Pytorch [30] που χρησιμοποιούνται έμμεσα μέσω άλλων βιβλιοθηκών επίσης παραλείπονται.

transformers

Η βιβλιοθήκη transformers παρέχει προηγμένα μοντέλα μηχανικής μάθησης για επεξεργασία φυσικής γλώσσας. Συγκεκριμένα, περιέχει προ-εκπαιδευμένα μοντέλα όπως το BERT [31], GPT [6] και πολλά άλλα που μπορούν να χρησιμοποιηθούν για διάφορες εργασίες, όπως μετάφραση, αναγνώριση ονομαστικών οντοτήτων, σύνθεση κειμένου και άλλες.

nltk

Το nltk είναι μία βιβλιοθήκη φυσικής γλώσσας για την Python. Παρέχει ποικίλα εργαλεία και δεδομένα για την επεξεργασία κειμένου, όπως τον διαχωρισμό λέξεων, τον υπολογισμό στατιστικών μέτρων κειμένου, την συντακτική ανάλυση και πολλά άλλα.

googlesearch

Η βιβλιοθήκη googlesearch παρέχει μια διεπαφή, για να πραγματοποιηθεί αυτοματοποιημένα αναζήτηση στον ιστό, χρησιμοποιώντας τη μηχανή αναζήτησης της Google. Υπάρχει δυνατότητα να γίνει χρήση της για αυτόματες αναζητήσεις σε διαφορετικές γλώσσες, με διαφορετικό αριθμό αποτελεσμάτων, με διαφορετικά κριτήρια, σε διαφορετικούς χρονισμούς.

urllib

Η βιβλιοθήκη urllib παρέχει εργαλεία για την ανάκτηση, τον χειρισμό και την αποστολή δεδομένων μέσω του διαδικτύου. Χρησιμοποιείται για την λήψη ιστοσελίδων, το άνοιγμα URLs, τις αιτήσεις HTTP και την επεξεργασία δεδομένων από τον ιστό.

sklearn

Η βιβλιοθήκη sklearn (ή scikit-learn) είναι μια από τις πιο δημοφιλείς βιβλιοθήκες μηχανικής μάθησης στην Python. Παρέχει εργαλεία για την εκπαίδευση και την εφαρμογή διάφορων αλγορίθμων μηχανικής μάθησης, όπως αποτελεσματικούς ταξινομητές, ανιχνευτές ανωμαλιών, μεθόδους συσταδοποίησης και πολλά άλλα. Επιπλέον, περιλαμβάνει εργαλεία για την προεπεξεργασία δεδομένων και την αξιολόγηση μοντέλων.

lxml

Η βιβλιοθήκη lxml είναι μια εξειδικευμένη βιβλιοθήκη για την προσπέλαση και τον χειρισμό XML και HTML δεδομένων. Παρέχει μια γρήγορη και ευέλικτη δυνατότητα ανάλυσης δομημένων εγγράφων, επιτρέποντας την ανάκτηση και την επεξεργασία δεδομένων από αρχεία XML και HTML με άνεση.

bs4

Η βιβλιοθήκη bs4 (ή Beautiful Soup 4) είναι ένα ισχυρό εργαλείο για την προσπέλαση και την εξαγωγή δεδομένων από ιστοσελίδες. Επιτρέπει την περιήγηση σε ένα XML ή HTML έγγραφο, τον εντοπισμό στοιχείων με βάση τις ετικέτες τους, τις κλάσεις τους, τα αναγνωριστικά τους και άλλα χαρακτηριστικά, και την εξαγωγή δεδομένων.

multiprocessing

Η βιβλιοθήκη multiprocessing παρέχει ένα ισχυρό πλαίσιο για παράλληλους υπολογισμούς και πολυεπεξεργασία. Επιτρέπει την αξιοποίηση πολλαπλών επεξεργαστών ή πυρήνων για την ταυτόχρονη εκτέλεση εργασιών, βελτιώνοντας έτσι την απόδοση και την αποδοτικότητα του κώδικά. Με την multiprocessing, δύναται η δημιουργία και η διαχείριση διεργασιών, νημάτων και ουρών εργασιών, επιτρέποντας τον σχεδιασμό και την υλοποίηση επεκτάσιμων και ευέλικτων εφαρμογών.

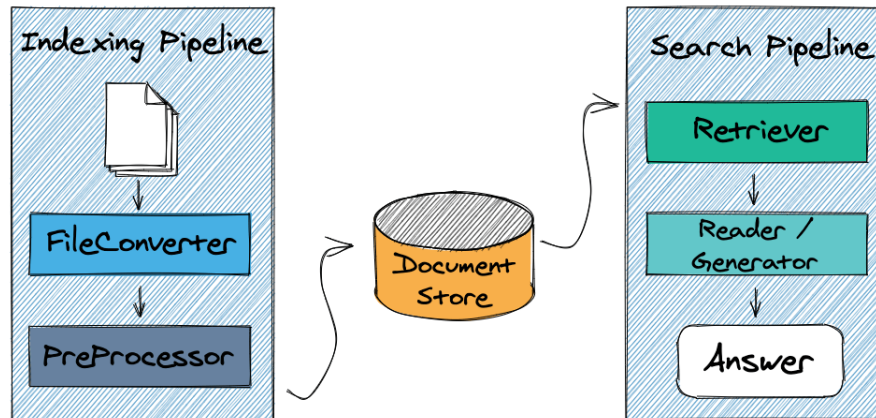
matplotlib

Η πλειονότητα των διαγραμμάτων έγινε με το matplotlib.pyplot. Η matplotlib είναι μια βιβλιοθήκη στην python για τη δημιουργία γραφικών και διαγραμμάτων. Χρησιμοποιείται ευρέως για οπτικοποίηση δεδομένων και παρέχει πολλές επιλογές προσαρμογής. Είναι ενσωματωμένη με την python και εύκολη στη χρήση για δημιουργία διαγραμμάτων. Η matplotlib.pyplot είναι μια διεπαφή βασισμένη στην matplotlib. Παρέχει έναν έμμεσο, παρόμοιο με το MATLAB, τρόπο σχεδίασης. Παράγει γραφικές αναπαραστάσεις και ενεργεί ως διαχειριστής της διεπαφής των γραφικών αναπαραστάσεων.

3.2 Haystack

Το Haystack [32] είναι ένα πλαίσιο ανοικτού κώδικα για τη δημιουργία συστημάτων αναζήτησης που λειτουργούν έξυπνα σε μεγάλες συλλογές εγγράφων. Οι πρόσφατες εξελίξεις στον τομέα του NLP επέτρεψαν την εφαρμογή της απάντησης ερωτήσεων, της ανάκτησης και της σύνοψης σε πραγματικές συνθήκες, και το Haystack έχει σχεδιαστεί για να αποτελέσει τη γέφυρα μεταξύ της έρευνας και της βιομηχανίας. Το Haystack επιτρέπει την χρήση βάσεων δεδομένων, διαφορετικών ήδη εκπαιδευμένων μοντέλων, διαφορετικών εξαρτημάτων.

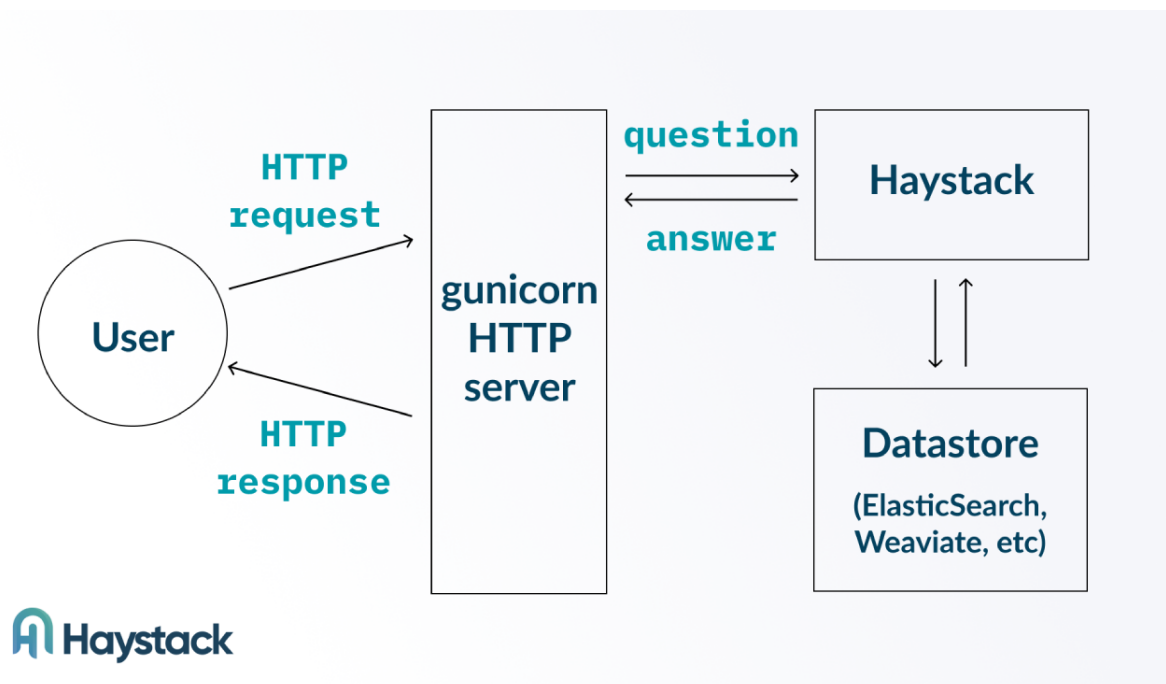
Το Haystack απαρτίζεται από 3 διαφορετικά επίπεδα (Σχήμα 3.1). Τους κόμβους, τα pipelines και το REST API. Οι κόμβοι είναι τα δομικά στοιχεία ενός pipeline και μπορούν να χρησιμοποιηθούν αυτόνομα και με διαφορετικούς συνδυασμούς. Τα pipelines είναι μια ακολουθία από συνδεδεμένα στοιχεία τα οποία



Σχήμα 3.1: Τυπική Μορφή ενός συστήματος στο Haystack

εκτελούν μια διεργασία. Για παράδειγμα, μια αλυσίδα από έναν Reader και έναν Retriever χτίζουν ένα pipeline ερωτοαπάντησης.

Το REST API είναι η αρχιτεκτονική που χρησιμοποιείται για τη διαχείριση αιτημάτων σε διαδικτυακές εφαρμογές (Σχήμα 3.2).



Σχήμα 3.2: Το REST API του Haystack

Στην πιο απλή αρχιτεκτονική, αυτή του παραπάνω σχήματος, διακρίνονται 6 διαφορετικά επίπεδα. Στο πρώτο επίπεδο ένας μετατροπέας αρχείων οδηγεί τα διαθέσιμα αρχεία σε έναν προ-επεξεργαστή. Αυτά αποθηκεύονται με τη νέα τους μορφή σε ένα Document Store. Το Document Store είναι μια βάση δεδομένων για αποθήκευση και ανάκτηση εγγράφων και μεταδεδομένων. Συνήθως, η προέλευση των εγγράφων είναι το διαδίκτυο ή κάποια άλλη εξωτερική πηγή. Χαρακτηρίζονται από αποτελεσματική προσπέλαση και γρήγορη ανάκτηση της πληροφορίας. Στην παρούσα διπλωματική εργασία γίνεται χρήση του Document Store Elasticsearch.

Στη συνέχεια του μοντέλου ένα Retriever ανακτά τα αρχεία από τη βάση δεδομένων και τα προωθεί στον Reader. Το Retriever εκτελεί ανάκτηση εγγράφων σαρώνοντας ένα Document Store και επιστρέφοντας ένα σύνολο υποψήφιων εγγράφων που σχετίζονται με το ερώτημα. Ο Reader λαμβάνει μια ερώτηση και ένα σύνολο εγγράφων ως είσοδο και επιστρέφει μια απάντηση επιλέγοντας ένα πλαίσιο κειμένου μέσα στα έγγραφα. Οι Readers χρησιμοποιούν μοντέλα, για να πραγματοποιούν ερωτοαπαντήσεις. Στην παρούσα διπλωματική εργασία γίνεται χρήση του Reader FARMReader [16] και του TransformersReader [16].

Elasticsearch

Το Elasticsearch είναι μια κατανεμημένη, δωρεάν και ανοικτή μηχανή αναζήτησης και ανάλυσης για όλους τους τύπους δεδομένων, συμπεριλαμβανομένων των κειμενικών, αριθμητικών, γεωχωρικών, δομημένων και αδόμητων δεδομένων [33].

Η παροχή απλού REST APIs, η κατανεμημένη φύση του, η ταχύτητα και η επεκτασιμότητα το κάνουν το κεντρικό στοιχείο της Elastic Stack, ενός συνόλου δωρεάν και ανοικτών εργαλείων για την εισαγωγή, τον εμπλουτισμό, την αποθήκευση, την ανάλυση και την οπτικοποίηση δεδομένων [33].

Η ταχύτητα και η επεκτασιμότητα του Elasticsearch και η ικανότητά του να ευρετηριάζει πολλούς τύπους περιεχομένου σημαίνουν ότι μπορεί να χρησιμοποιηθεί για πολλές περιπτώσεις χρήσης. Κάποιες από αυτές είναι η Αναζήτηση εφαρμογών, η Αναζήτηση ιστοτόπων, η Επιχειρησιακή αναζήτηση, η Καταγραφή και ανάλυση αρχείων καταγραφής, η Μετρήσεις υποδομής και παρακολούθηση δοχείων, η Παρακολούθηση επιδόσεων εφαρμογών, η Ανάλυση και οπτικοποίηση γεωχωρικών δεδομένων, η Ανάλυση ασφάλειας, και η Επιχειρηματική ανάλυση [33].

Η χρήση βασίζεται σε μερικές βασικές αρχές. Τα ακατέργαστα δεδομένα εισρέουν στο Elasticsearch από διάφορες πηγές, συμπεριλαμβανομένων των αρχείων καταγραφής, των μετρικών του συστήματος και των διαδικτυακών εφαρμογών. Η εισαγωγή δεδομένων είναι η διαδικασία με την οποία αυτά τα ακατέργαστα δεδομένα αναλύονται, κανονικοποιούνται και εμπλουτίζονται προτού ευρετηριαστούν στο Elasticsearch. Αφού ευρετηριαστούν στο Elasticsearch, οι χρήστες μπορούν να εκτελούν σύνθετα ερωτήματα στα δεδομένα τους και να χρησιμοποιούν συσσωρεύσεις, για να ανακτούν σύνθετες περιλήψεις των δεδομένων τους [33].

Ένα ευρετήριο του Elasticsearch είναι μια συλλογή εγγράφων που σχετίζονται μεταξύ τους. Το Elasticsearch αποθηκεύει δεδομένα ως έγγραφα JSON. Κάθε έγγραφο συσχετίζει ένα σύνολο κλειδίων (ονόματα πεδίων ή ιδιοτήτων) με τις αντίστοιχες τιμές τους (συμβολοσειρές, αριθμούς, Booleans, ημερομηνίες, πίνακες τιμών, γεωγραφικές θέσεις ή άλλους τύπους δεδομένων).

Το Elasticsearch χρησιμοποιεί μια δομή δεδομένων που ονομάζεται ανεστραμμένο ευρετήριο Inverted Indices, η οποία έχει σχεδιαστεί για να επιτρέπει πολύ γρήγορες αναζητήσεις πλήρους κειμένου. Ένα ανεστραμμένο ευρετήριο απαριθμεί κάθε μοναδική λέξη που εμφανίζεται σε οποιοδήποτε έγγραφο και προσδιορίζει όλα τα έγγραφα στα οποία εμφανίζεται κάθε λέξη.

Κατά τη διάρκεια της διαδικασίας ευρετηρίασης, το Elasticsearch αποθηκεύει έγγραφα και δημιουργεί ένα ανεστραμμένο ευρετήριο, για να καταστήσει τα δεδομένα των εγγράφων αναζητήσιμα σε σχεδόν πραγματικό χρόνο. Η ευρετηρίαση ξεκινά με το API ευρετηρίου, μέσω του οποίου μπορείτε να προσθέσετε ή να ενημερώσετε ένα έγγραφο JSON σε ένα συγκεκριμένο ευρετήριο [33].

Στην παρούσα διπλωματική εργασία το Elasticsearch χρησιμοποιήθηκε για ευρετηρίαση, αποθήκευση και προσπέλαση πληροφορίας.

Readers

Οι δύο κυριότεροι Readers είναι ο FARMReader και ο TransformersReader. Εκτός από τα βάρη των μοντέλων, οι Haystack Readers περιέχουν όλα τα στοιχεία που συναντώνται σε end-to-end συστήματα ερωτοαπαντήσεων. Ένα τέτοιο σύστημα περιλαμβάνει την κωδικοποίηση, τον υπολογισμό της ενσωμάτωσης, την πρόβλεψη της έκτασης και τη συγκέντρωση των υποψηφίων απαντήσεων. Οι βιβλιοθήκες FARM και Transformers χειρίζονται τα βάρη με τον ίδιο τρόπο, αλλά τα QA pipelines τους διαφέρουν.

- Ο TransformersReader μερικές φορές προβλέπει το ίδιο span δύο φορές. Ο FARMReader αφαιρεί τα διπλά spans.

- Ο FARMReader χρησιμοποιεί επί του παρόντος τους tokenizers από τη βιβλιοθήκη Hugging Face Transformers. Ο TransformersReader χρησιμοποιεί τους tokenizers από τη βιβλιοθήκη Hugging Face Tokenizers.
- Ο TransformersReader κανονικοποιεί τα logits αρχής και τέλους ανά πέρασμα, και τα πολλαπλασιάζει. Ο FARMReader τα αθροίζει και δεν τα κανονικοποιεί.

3.3 Flask

Το Flask είναι ένα μικροπλαίσιο ιστού γραμμένο σε Python. Χαρακτηρίζεται ως μικροπλαίσιο, επειδή δεν απαιτεί συγκεκριμένα εργαλεία ή βιβλιοθήκες. Δεν έχει επίπεδο βάσεων δεδομένων, επικύρωση φόρμας ή άλλα στοιχεία όπου προϋπάρχουσες βιβλιοθήκες τρίτων παρέχουν κοινές λειτουργίες. Ωστόσο, το Flask υποστηρίζει επεκτάσεις που μπορούν να προσθέσουν χαρακτηριστικά εφαρμογών σαν να είχαν υλοποιηθεί στο ίδιο το Flask. Υπάρχουν επεκτάσεις για χαρτογράφους αντικειμενοστρεφούς απεικόνισης, επικύρωση φόρμας, χειρισμό μεταφόρτωσης, διάφορες ανοικτές τεχνολογίες αυθεντικοποίησης και διάφορα κοινά εργαλεία που σχετίζονται με το πλαίσιο.

Το Flask χρησιμοποιεί τη μηχανή προτύπων Jinja για τη δυναμική δημιουργία σελίδων HTML, χρησιμοποιώντας οικείες έννοιες της Python, όπως μεταβλητές, βρόχους, λίστες και ούτω καθεξής. Επομένως, κατά τη χρήση αυτού του μικροπλαισίου είναι σύνηθες να υλοποιείται η διεπαφή χρήστη (UI) με τη χρήση HTML και Javascript.

Στις εφαρμογές που χρησιμοποιούν το Flask περιλαμβάνονται δημοφιλείς πλατφόρμες με δισεκατομμύρια χρήστες, όπως το Pinterest και το LinkedIn.

Κεφάλαιο 4

Μεθοδολογία

Σκοπός αυτής της διπλωματικής εργασίας είναι η ανάπτυξη ενός ιδιωτικού δυναμικού ψηφιακού βοηθού ερωταπαντήσεων με προσωποποιημένο χαρακτήρα. Η δυναμικότητα του συστήματος επιτυγχάνεται με την αναζήτηση πληροφορίας στον παγκόσμιο ιστό. Η αναζήτηση πραγματοποιείται σε πραγματικό χρόνο έπειτα από το ερώτημα του χρήστη. Η ιδιωτικότητα επιτυγχάνεται μέσω της εξαγωγής της βάσης δεδομένων και την λειτουργία του συστήματος σε κατάσταση εκτός σύνδεσης (βλέπε ενότητα 4.5). Με τον ίδιο τρόπο ο χρήστης μπορεί να έχει μια προσωποποιημένη βάση δεδομένων που να την χρησιμοποιεί ως πηγή για τις απαντήσεις στα ερωτήματα που υποβάλλονται στο σύστημα.

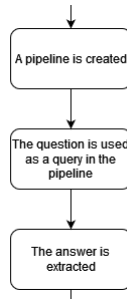
4.1 Αλγόριθμος

Ο αλγόριθμος που ακολουθείται στην παρούσα διπλωματική εργασία αποτελείται από 6 κόμβους απόφασης. Ανάλογα με τις επιλογές του χρήστη, τα αποτελέσματα που εντοπίζει ο αλγόριθμος, τα αποτελέσματα των μοντέλων ερωταπαντήσεων και τις επιλογές του προγραμματιστή, κάποια βήματα παραλείπονται ή επαναλαμβάνονται. Αιτία αυτών των επαναλήψεων η παραλείψεις μπορεί να είναι είτε η απουσία ικανοποιητικής απάντησης, πχ χαμηλό σκορ, είτε η επιλογή λειτουργίας γρήγορης απάντησης από τον χρήστη ή η παράλειψη κάποιου βήματος από τον προγραμματιστή για λόγους ταχύτητας. Κατά τον τερματισμό του αλγορίθμου επιστρέφονται οι πιθανές απαντήσεις στο ερώτημα του χρήστη. Το πλήθος τους εξαρτάται από τον αριθμό των βημάτων αλλά και την ποιότητα της πληροφορίας που χρησιμοποιήθηκε.

4.1.1 Χρήσιμοι Όροι

Για τη σωστή και λεπτομερή περιγραφή του αλγορίθμου χρειάζεται να οριστούν και επεξηγηθούν κάποιες έννοιες που χρησιμοποιούνται κατά την ανάλυσή του. Ως **ερώτημα** του χρήστη ορίζεται το query που εισάγει ο χρήστης στο κουτί της φόρμας που χρησιμοποιείται σαν είσοδος. Η εισαγωγή γίνεται μέσω του πληκτρολογίου και υποβάλλεται με το πάτημα του κουμπιού submit. Ως **πηγή** ορίζεται το αρχείο μορφής txt ή η μεταβλητή στα οποία περιέχονται οι πληροφορίες πάνω στις οποίες το μοντέλο ερωταπαντήσεων βασίζεται για να δώσει την απάντηση στο ερώτημα. Ως **απάντηση** ορίζεται το σύνολο των απαντήσεων, του σκορ τους, του συνδέσμου από όπου προέρχεται η πηγή τους αλλά και του πλαισίου στο οποίο εντοπίστηκαν. (answer, score, url και context). Ως **σκορ** μιας απάντησης ορίζεται ο δεκαδικός αριθμός μεταξύ του 0 και του 1 που επιστρέφεται από τον Retriever και δηλώνει την ποιότητα της απάντησης. 1 είναι το άριστο και 0 το χαμηλότερο σκορ. Ως **τιμή limit** ορίζεται αριθμητική τιμή μεταξύ του 0 και του 1, όπου απάντηση με σκορ πάνω από αυτήν θεωρείται αποδεκτή. Στο πρόγραμμα που εκτελέστηκε κατά το Κεφάλαιο 5, όπου έγιναν οι συγκρίσεις, η τιμή **limit** είναι το 0,9 αλλά θα μπορούσε να ποικίλλει σε διαφορετικά σημεία του αλγορίθμου. Ως **σημαντική λέξη** ορίζεται μια λέξη η οποία χρησιμοποιείται για την αξιολόγηση των αποθηκευμένων συνδέσμων του πρώτου αποτελέσματος. Μπορεί να είναι μια ή περισσότερες. Ως **κορυφαίο αποτέλεσμα** ορίζεται το πρώτο αποτέλεσμα της μηχανής αναζήτησης της Google. Ως **Διαδικασία Απάντησης** ορίζεται η δημιουργία ενός Pipeline, η εισαγωγή του ερωτήματος στο Pipeline και η εξαγωγή της απάντησης με πηγή

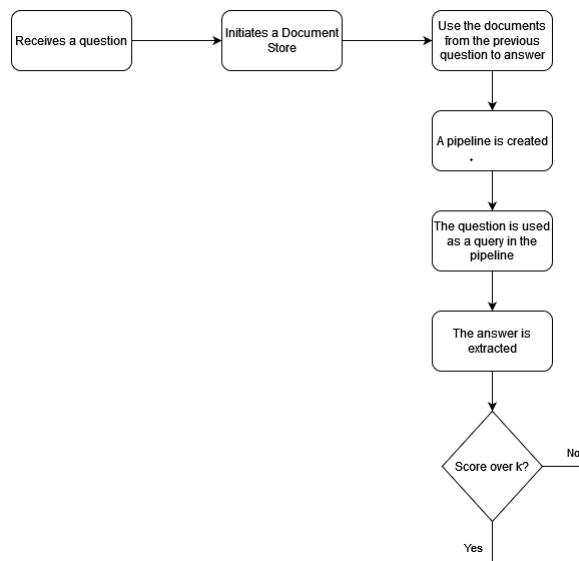
το περιεχόμενο του Document Store. Στο Σχήμα 4.1 φαίνεται η διαδικασία σε διάγραμμα ροής.



Σχήμα 4.1: Διαδικασία Απάντησης

4.1.2 Απάντηση χωρίς αναζήτηση

Η έναρξη του αλγορίθμου ορίζεται από τον χρήστη. Ο χρήστης εισάγει ένα ερώτημα στο κουτί της φόρμας, (text field στην HTML), που εμφανίζεται στην οθόνη του όταν φορτώσει η σελίδα. Με το πάτημα του κουμπιού submit ξεκινάει αυτόματα από το backend η έναρξη ενός Document Store, το οποίο θα χρησιμοποιηθεί για την αποθήκευση των πηγών. Ανεξάρτητα εάν έχει γίνει ήδη άλλο ερώτημα νωρίτερα από τον ίδιο χρήστη, θα γίνει μια προσπάθεια να απαντηθεί το νέο ερώτημα με την πηγή προηγούμενων ερωτημάτων. Σκοπός αυτής της κίνησης είναι να δοθεί μια γρήγορη απάντηση επωφελοόμενοι πιθανό κοινό πλαίσιο στο νέο και το παλιό ερώτημα, διότι η αναζήτηση πηγών στο διαδίκτυο αποτελεί ένα ιδιαίτερα χρονοβόρο κομμάτι του αλγορίθμου. Κάθε απάντηση ελέγχεται ως προς την ποιότητά της. Σε περίπτωση που το σκορ της απάντησης είναι μεγαλύτερο της τιμής *limit* επιστρέφεται η απάντηση, εμφανίζεται στην οθόνη και τερματίζει ο αλγόριθμος. Αντιθέτως, μια απάντηση με σκορ κατώτερο του ορίου αναγκάζει τον αλγόριθμο να συνεχίσει (Σχήμα 4.2).

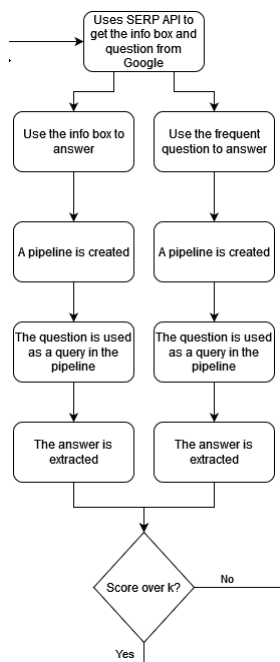


Σχήμα 4.2: Απάντηση χωρίς αναζήτηση

4.1.3 Απάντηση με αναζήτηση με το SERP API

Στο επόμενο βήμα γίνεται χρήση του SERP API για την λήψη του info box και των συχνών ερωτήσεων που παρέχει η μηχανή αναζήτησης της Google. Το νέο pipeline χρησιμοποιεί ως πηγή τις νέες αυτές πληροφορίες

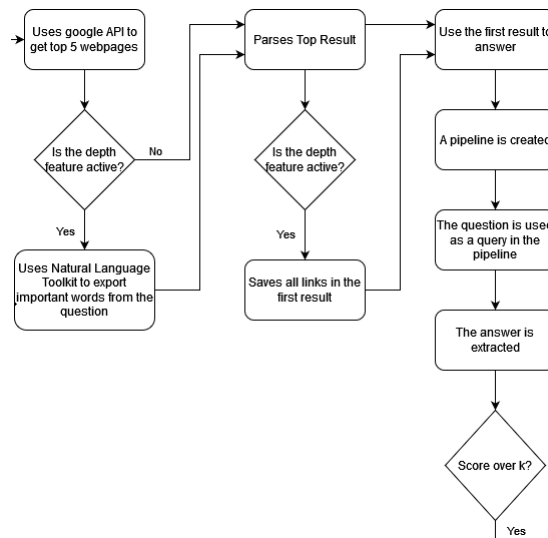
που έχουν αποθηκευτεί σε ένα νέο έγγραφο. Πραγματοποιείται έλεγχος της ποιότητας της απάντησης και λαμβάνεται μια απόφαση όπως παραπάνω (Σχήμα 4.3).



Σχήμα 4.3: Απάντηση με το SERP API

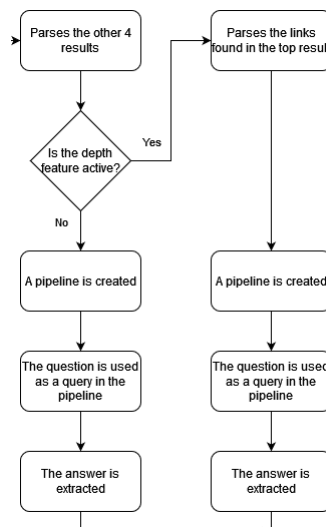
4.1.4 Απάντηση με αναζήτηση στο διαδίκτυο

Σε περίπτωση μη-ικανοποιητικής απάντησης γίνεται χρήση του API της Google. Ως όρισμα δίνεται το ερώτημα του χρήστη. Αποθηκεύονται τα κορυφαία 5 μοναδικά αποτελέσματα. Σε περίπτωση που η λειτουργία βάθους είναι ενεργή γίνεται χρήση της βιβλιοθήκης NLT για την εξαγωγή των σημαντικών λέξεων της ερώτησης. Ακολούθως, γίνεται προσπέλαση του κορυφαίου αποτελέσματος και χρησιμοποιείται το περιεχόμενό του, έπειτα από στοιχειώδη προεπεξεργασία, ως πηγή. Σε περίπτωση που η λειτουργία βάθους είναι ενεργή αποθηκεύονται όλοι οι σύνδεσμοι που εντοπίζονται στην πηγή. Το νέο pipeline επιστρέφει ένα αποτέλεσμα το οποίο, ομοίως με πριν, αξιολογείται (Σχήμα 4.4).



Σχήμα 4.4: Απάντηση με το πρώτο αποτέλεσμα της αναζήτησης στο διαδίκτυο

Σε περίπτωση μη αποδεκτού αποτελέσματος προσπελάζονται και οι υπόλοιπες 4 ιστοσελίδες, και το περιεχόμενό τους χρησιμοποιείται ως πηγή. Σε περίπτωση που η λειτουργία βάθους είναι ενεργή αξιολογούνται οι αποθηκευμένοι σύνδεσμοι που βρέθηκαν στην πρώτη σελίδα. Η αξιολόγηση γίνεται με τη βοήθεια των σημαντικών λέξεων. Οι σύνδεσμοι που θα θεωρηθούν αξιόλογοι θα προσπελαστούν και το περιεχόμενό τους θα προστεθεί στην πηγή. Το νέο πλέον pipeline θα δώσει μια τελική απάντηση. Ο αλγόριθμος τερματίζεται (Σχήμα 4.5).



Σχήμα 4.5: Απάντηση με όλα τα αποτελέσματα της αναζήτησης στο διαδίκτυο

Τα βήματα του αλγορίθμου σε κωδικοποιημένη μορφή είναι τα ακόλουθα :

1. Λήψη του ερωτήματος του χρήστη.
2. Έναρξη ενός Document Store.
3. Χρήση του Document Store πιθανών προηγούμενων ερωτημάτων για την απάντηση της ερώτησης.
4. Δημιουργία pipeline.

5. Εισαγωγή του ερωτήματος στο pipeline.
6. Εξαγωγή απάντησης με πηγή το περιεχόμενο του Document Store.
7. Έλεγχος σκορ απάντησης. Σε περίπτωση που είναι μεγαλύτερο της τιμής *limit*, επιστροφή της απάντησης και τερματισμός.
8. Χρήση του SERP API για τη λήψη του info box και των συχνών ερωτήσεων από τη μηχανή αναζήτησης της Google.
9. Δημιουργία pipeline, εισαγωγή του ερωτήματος στο pipeline και εξαγωγή της απάντησης χρησιμοποιώντας το info box και τις συχνές ερωτήσεις ως πηγή.
10. Έλεγχος σκορ απάντησης. Σε περίπτωση που είναι μεγαλύτερο της τιμής *limit*, επιστροφή της απάντησης και τερματισμός.
11. Χρήση του API της Google για τη λήψη των 5 κορυφαίων ιστοσελίδων για τη συγκεκριμένη ερώτηση.
12. Σε περίπτωση που η λειτουργία βάθους δεν είναι ενεργή, μεταπήδηση στο βήμα 14.
13. Χρήση της βιβλιοθήκης NLT για εξαγωγή των σημαντικών λέξεων της ερώτησης.
14. Προσπέλαση του κορυφαίου αποτελέσματος της μηχανής αναζήτησης.
15. Σε περίπτωση που η λειτουργία βάθους είναι ενεργή, αποθήκευση όλων των συνδέσμων που θα βρεθούν στο κορυφαίο αποτέλεσμα.
16. Δημιουργία pipeline.
17. Εισαγωγή του ερωτήματος στο pipeline.
18. Εξαγωγή απάντησης με πηγή το κορυφαίο αποτέλεσμα.
19. Έλεγχος σκορ απάντησης. Σε περίπτωση που είναι μεγαλύτερο της τιμής *limit*, επιστροφή της απάντησης και τερματισμός.
20. Προσπέλαση των υπόλοιπων τεσσάρων αποτελεσμάτων της μηχανής αναζήτησης.
21. Σε περίπτωση που η λειτουργία βάθους δεν είναι ενεργή, μεταπήδηση στο βήμα 25.
22. Προσπέλαση των αποθηκευμένων ιστοσελίδων του βήματος 15.
23. Αξιολόγηση αυτών των ιστοσελίδων με τη χρήση των σημαντικών λέξεων και απόρριψη των μη αποδεκτών.
24. Δημιουργία pipeline.
25. Εισαγωγή του ερωτήματος στο pipeline.
26. Εξαγωγή απάντησης με πηγή τα τέσσερα (ή περισσότερα αν η λειτουργία βάθους είναι ενεργή) υπόλοιπα αποτελέσματα της μηχανής αναζήτησης.
27. Επιστροφή της απάντησης.

4.1.5 Πρώτες παρατηρήσεις

Σε αυτό το στάδιο πραγματοποιήθηκαν κάποιες πρώτες παρατηρήσεις και αξιολογήσεις. Αρχικά, το SERP API δίνει πολύ σωστά αποτελέσματα μόνο στις περιπτώσεις όπου το ερώτημα είναι γενικής φύσεως. Για παράδειγμα, στο ερώτημα "What is bitcoin?" το SERP API επιστρέφει τη συχνή ερώτηση "What is a Bitcoin and how is it made?" (Σχήμα 4.6), της οποίας η απάντηση είναι ακριβώς το επιθυμητό αποτέλεσμα. Πιο δύσκολα ερωτήματα αποτυγχάνουν να απαντηθούν. Ωστόσο, σε αυτές τις περιπτώσεις που το αποτέλεσμα είναι σωστό, η απάντηση είναι υψηλής ποιότητας, υψηλού σκορ, αλλά υστερεί στο πλαίσιο το οποίο επιστρέφεται στον χρήστη.

When computers on the network verify and process transactions, new bitcoins are created, or mined. These networked computers, or miners, process the transaction in exchange for a payment in Bitcoin. Bitcoin is powered by blockchain, which is the technology that powers many cryptocurrencies. 27 Mar 2023

Σχήμα 4.6: "What is bitcoin?" Συχνή Ερώτηση

Επιπρόσθετα, η λειτουργία βάθους έδειξε πως είναι πολύ χρονοβόρα, κοστοβόρα σε υπολογιστικούς πόρους και σπανίως προσφέρει ποιότητα στην απάντηση. Δεν προσδίδει πληροφορία ούτε ως προς το σκορ, ούτε ως προς την υποκειμενική ποιότητα της απάντησης κατά την γνώμη του γράφοντα, ούτε ως προς το περιεχόμενο στο οποίο βρέθηκε. Στη συντριπτική πλειοψηφία των περιπτώσεων, οι απαντήσεις που δίνονται μετά τη λειτουργία βάθους έχουν χαμηλότερο σκορ, ή ταυτίζονται με αυτές των προηγούμενων βημάτων. Μια πιθανή χρησιμότητα θεωρήθηκε σε περιπτώσεις που η απάντηση πρέπει να δοθεί με πηγή μια συγκεκριμένη ιστοσελίδα, οπότε και θα πρέπει να αξιοποιηθεί σε βάθος το υλικό της. Τέλος, στις περισσότερες περιπτώσεις το τελευταίο βήμα, ακόμα και χωρίς την λειτουργία βάθους, κρίνεται περιττό καθώς ήδη έχει δοθεί ικανοποιητική απάντηση.

4.2 Επιμέρους διαδικασίες

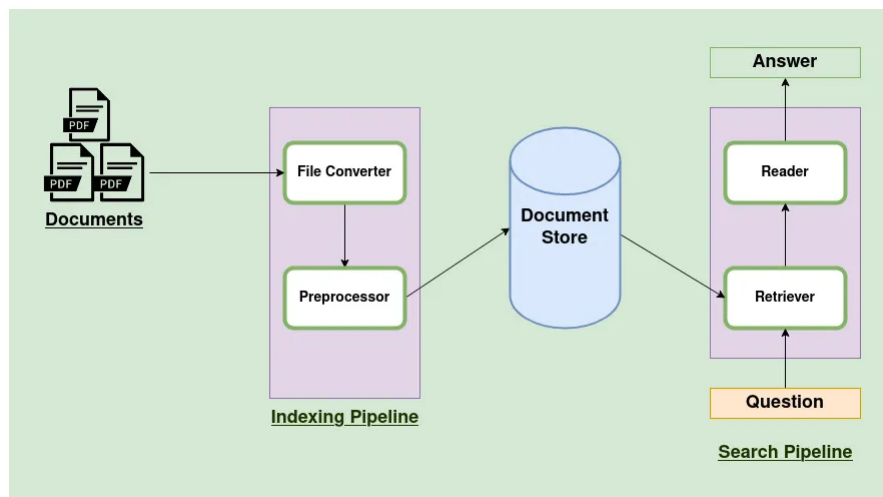
4.2.1 Ορίσματα Χρήστη

Ο χρήστης μέσω της διεπαφής δίνει κάποια ορίσματα. Τα question, language, website, enable quick answer και detailed answer. Στην δοκιμαστική λειτουργία που είναι υλοποιημένη, υπάρχουν και κάποια ορίσματα που κανονικά είναι διαθέσιμα μόνο κατά το setup του προγράμματος. Το κυριότερο όρισμα είναι το ερώτημα του χρήστη (question). Αυτό το όρισμα δεν μπορεί να είναι κενό και θα αποτελέσει την είσοδο σε κάθε pipeline κατά την διάρκεια όλων των βημάτων του αλγορίθμου. Το όρισμα της γλώσσας (language) καθορίζει τη γλώσσα στην οποία θα γίνει αναζήτηση της απάντησης στο ερώτημα του χρήστη. Το όρισμα ιστοσελίδας (website) καθορίζει σε ποιο επίπεδο θα γίνει η αναζήτηση πηγών για την απάντηση του ερωτήματος. Το default όρισμα είναι όλο το διαδίκτυο και χρησιμοποιείται τότε η μηχανή αναζήτησης της Google. Διαφορετικά, ο χρήστης μπορεί να επιλέξει κάποια άλλη ιστοσελίδα. Σε μια τέτοια περίπτωση ο αλγόριθμος θα αναζητήσει την απάντηση μόνο στον συγκεκριμένο ιστότοπο. Αυτή η λειτουργία προσπαθεί να προσομοιώσει πιθανή μεταφορά της εφαρμογής σε plug-in, το οποίο θα μπορεί να χρησιμοποιηθεί σε οποιαδήποτε ιστοσελίδα για αναζήτηση μόνο εντός αυτής. Το όρισμα της γρήγορης απάντησης (enable quick answer) όταν είναι επιλεγμένο, είναι boolean τιμή, επιτρέπει την αναζήτηση μέσω του SERP API. Η ύπαρξη αυτού του ορίσματος οφείλεται στο γεγονός πως το συγκεκριμένο API δεν είναι δωρεάν και η λειτουργία δωρεάν δοκιμής επιτρέπει ορισμένο αριθμό καλεσμάτων τον μήνα. Το όρισμα λεπτομερούς απάντησης είναι για σκοπούς testing. Με αυτήν την επιλογή ενεργοποιημένη επιστρέφονται όλες οι απαντήσεις, ανεξαρτήτως σκορ και καταλληλότητας.

4.2.2 Pipeline

Το pipeline, στα Ελληνικά αγωγός μηχανικής μάθησης, είναι η end-to-end κατασκευή που ενορχηστρώνει τη ροή δεδομένων σε ένα μοντέλο μηχανικής μάθησης (ή ένα σύνολο πολλαπλών μοντέλων) και την έξοδο από αυτό. Περιλαμβάνει την είσοδο ακατέργαστων δεδομένων, τα χαρακτηριστικά, τις εξόδους, το μοντέλο μηχανικής μάθησης και τις παραμέτρους του μοντέλου, καθώς και τις εξόδους πρόβλεψης.

Το pipeline στον συγκεκριμένο αλγόριθμο έχει 5 κόμβους. Αποτελείται από έναν Text Converter (στο Σχήμα 4.7 αναφέρεται ως File Converter), έναν Pre Processor, ένα Document Store, έναν Retriever και έναν Reader. Η είσοδος του Text Converter είναι το είδος του αρχείου που θα χρησιμοποιηθεί ως πηγή, η είσοδος του Pre Processor είναι ο Text Converter, η είσοδος του Document Store είναι ο Pre Processor, η είσοδος του Retriever είναι το query, δηλαδή το ερώτημα του χρήστη μαζί και οι πληροφορίες που είναι αποθηκευμένες στο Document Store, και η είσοδος του Reader είναι ο Retriever (Σχήμα 4.7).



Σχήμα 4.7: Δομή ενός pipeline [34]

Ο Retriever που χρησιμοποιείται είναι ο BM25Retriever, ο Reader που χρησιμοποιείται είναι ο TransformersReader και το μοντέλο είναι το deepset/minilm-uncased-squad2.

Οι δυο Readers που χρησιμοποιήθηκαν είναι ο TransformersReader και ο FARMReader. Ο TransformersReader αποδείχτηκε γρηγορότερος και ελαφρώς πιο αδύναμος από τον FARMReader. Αυτή η αδυναμία δεν θεωρήθηκε αρκετή για την απόρριψή του.

Τα τρία μοντέλα για την Αγγλική γλώσσα που χρησιμοποιήθηκαν είναι τα minilm, roberta και xxlarge-v1. Το minilm είναι μοντέλο αυτοπροσοχής για συμπίεση προ-εκπαιδευμένων μετασχηματιστών ανεξαρτήτως μοντέλου [35]. Το roberta είναι ένα μοντέλο βελτιστοποιημένης προεκπαιδευμένης προσέγγισης του BERT [36]. Το xxlargev1 είναι ένα μοντέλο τύπου BERT για αυτοεποπτευόμενη εκμάθηση γλωσσικών αναπαραστάσεων [37]. Τα τρία πολυγλωσσικά μοντέλα που χρησιμοποιήθηκαν είναι τα xlm roberta, deBERTa και gelectra.

Το μοντέλο minilm αποδείχθηκε και πιο γρήγορο και πιο ακριβές από το roberta, όπως άλλωστε έλεγαν και οι προδιαγραφές. Από την άλλη το μοντέλο xxlargev1 παρουσίασε εξαιρετικά αποτελέσματα αλλά πάρα πολύ χαμηλές ταχύτητες.

4.2.3 Σημαντικές Λέξεις

Σε αυτήν την εργασία εφαρμόστηκαν δύο μέθοδοι για την εξαγωγή των σημαντικών λέξεων. Η πρώτη μέθοδος χρησιμοποιεί το μοντέλο BERT, για να αξιολογήσει τις λέξεις του ερωτήματος του χρήστη. Μέσω κατάλληλης βιβλιοθήκης χρησιμοποιούνται οι μετασχηματιστές του μοντέλου και εξάγονται κάποιες λέξεις με βάση το σκορ αυτοπροσοχής τους. Ωστόσο, αυτή η μέθοδος είναι ιδιαίτερα χρονοβόρα μειώνοντας σημαντικά την ταχύτητα του αλγορίθμου. Μια απλούστερη μέθοδος είναι η εξαγωγή των σημαντικών λέξεων κρίνοντας το μέρος του λόγου το οποίο είναι η λέξη. Έτσι, ως σημαντικές λέξεις επιλέγονται όλα τα ουσιαστικά και τα επίθετα της ερώτησης του χρήστη.

4.2.4 Αξιολόγηση Ιστοσελίδων Βάθους

Τρεις είναι οι κύριες προσεγγίσεις που υλοποιούν τη σύγκριση ομοιότητας μεταξύ μικρών κειμένων. Οι 3 μεθοδολογίες είναι word-to-word based, vector based και structured based.

Η word-to-word βρίσκει ομοιότητα μεταξύ λέξεων. Η ομοιότητα μεταξύ λέξεων είναι το μήκος του πιο κοντινού path μεταξύ των δυο λέξεων σε έναν γράφο με ολόκληρο το corpus. Αντίστοιχα η ομοιότητα μεταξύ λέξεων είναι ο ημιτονοειδής συντελεστής, cosine coefficient, $S_s = \frac{s_1 * s_2}{|s_1| * |s_2|}$. Αυτή η μέθοδος χρησιμοποιήθηκε για το χτίσιμο των δέντρων μέσω των λημμάτων της wikipedia [38].

Η vector προσέγγιση χρησιμοποιεί word embedding, δηλαδή μια διανυσματική απεικόνιση των λέξεων επομένως και των προτάσεων. Τα διανύσματα μέσω DL μοντέλων αποτυπώνουν σημασιολογικά χαρακτηριστικά των λέξεων [39]. Παράλληλα χρησιμοποιούνται η Συχνότητα Όρου και η Αντίστροφη Συχνότητα Όρου (TF-IDF), που χρησιμοποιείται και στην παρούσα εργασία, για περαιτέρω αποτύπωση της σημασίας των λέξεων.

Ενδιαφέρον παρουσιάζει και η δομική προσέγγιση, η οποία εστιάζει κυρίως στη γραμματική δομή των προτάσεων [40]. Στην παρούσα εργασία θεωρήθηκε μη εφαρμόσιμη.

Στη σύγκριση μεγαλύτερων κειμένων χρησιμοποιούνται αντίστοιχες τεχνικές. Το count vectorizer είναι απλώς η απαρίθμηση των λέξεων που δίνει έμφαση στις επαναλαμβανόμενες λέξεις αλλά όχι τόσο στις σχετικές. Αυτού του τύπου οι τεχνικές ονομάζονται Bag-Of-Words (BOW), αλλά αδυνατούν να αντιληφθούν τα semantics [41] κάτι που τις κάνει λιγότερο χρήσιμες για την παρούσα εργασία.

Επιπλέον μέθοδοι που χρησιμοποιούνται είναι ένας συνδυασμός tf-idf + Latent Semantic Analysis (LSA) και το doc2vec. Το LSA είναι μίξη του document-term πίνακα και του SVD (Singular Value Decomposition). Ο document-term πίνακας είναι ένας πίνακας συχνότητας εμφάνισης λέξεων και το SVD είναι η κακονικοποίηση του χώρου με τη βοήθεια ιδιοτιμών. Το doc2vec από την άλλη είναι η μέθοδος word2vec αλλά αναγόμενη σε έγγραφα.

Μετρικές για τη σχέση λέξεων θεωρούνται:

1. Fuzzywuzzy, συγκρίνει τις δυο συμβολοσειρές και υπολογίζει την απόσταση μεταξύ τους που απαιτείται ώστε αν τροποποιηθούν να ταυτίζονται
2. Jaccard Similarity $J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$
3. Cosine similarity

$$\text{cosine similarity} = S_c(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

4. Pearson Correlation Coefficient

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Στην προσέγγιση αυτή, χρησιμοποιήθηκαν 3 μέθοδοι για την αξιολόγηση και επιλογή των κατάλληλων Ιστοσελίδων Βάθους. Και οι 3 χρησιμοποιούν τις σημαντικές λέξεις. Η μια μέθοδος, η πιο απλή, απλώς θεωρεί μια ιστοσελίδα αποδεκτή αν μια από τις σημαντικές λέξεις είναι παρούσα στο κείμενο που εξάγεται από κάθε ιστοσελίδα/σύνδεσμο. Η δεύτερη μέθοδος χρησιμοποιεί τη Συχνότητα Όρου και τη Αντίστροφη Συχνότητα Όρου (TF-IDF). Επιστρέφει μια μετρική όπου και είναι και το σκορ αξιολόγησης της ιστοσελίδας. Η τρίτη μέθοδος χρησιμοποιεί τα μεταδεδομένα της ιστοσελίδας και αρκεί ο εντοπισμός μιας από των σημαντικών λέξεων. Όπως είναι προφανές, η τελευταία μέθοδος είναι η γρηγορότερη αλλά και με τη μικρότερη επιτυχία. Αυτό οφείλεται στο γεγονός πως χρειάζεται να γίνει λήψη μόνο των μεταδεδομένων, Metadata, και όχι της πλήρους ιστοσελίδας. Ταυτόχρονα όμως, προϋποθέτει πως τα μεταδεδομένα είναι πλούσια και λεπτομερή. Η TF-IDF μέθοδος είναι πιο αργή, αλλά έχει την καλύτερη ακρίβεια και επιστρέφει τη μέγιστη πληροφορία. Μάλιστα αυτή προτείνεται και από τη βιβλιογραφία.

4.3 Μοντέλα

Το σύστημα που αναπτύχθηκε στην παρούσα εργασία αποτελεί παράδειγμα συστήματος Ερωταπαντήσεων. Το σύστημα αυτό δέχεται ως είσοδο ορισμένο κείμενο (ερώτημα), και από αυτό με τη βοήθεια ενός μοντέλου παράγει έναν αριθμό απαντήσεων, χωρίς να έχουν ως συμπληρωματική είσοδο τις απαντήσεις των ερωτήσεων αυτών. Κατά την εκπόνηση αυτής της διπλωματικής εργασίας έγινε χρήση 6 διαφορετικών μοντέλων που παρουσιάζονται παρακάτω. Η επιλογή πολλών μοντέλων έγινε με το σκοπό την αναζήτηση και εύρεση αυτού που ταιριάζει περισσότερο στην συγκεκριμένη εφαρμογή.

Μοντέλα

Με τον Reader FARMReader χρησιμοποιούνται 6 μοντέλα. Τα 3 πρώτα είναι εκπαιδευμένα πάνω στην Αγγλική γλώσσα. Τα 3 τελευταία είναι multi-lingual:

- deepset/roberta-base-squad2 [10]: Το μοντέλο RoBERTa βασίζεται στο BERT και τροποποιεί βασικές υπερπαραμέτρους, αφαιρώντας τον στόχο προ-εκπαίδευσης της επόμενης πρότασης και εκπαιδευόμενος με πολύ μεγαλύτερες μίνι-παρτίδες και ρυθμούς μάθησης. Η ανάπτυξή του προέκυψε από τη διαπίστωση ότι το BERT ήταν σημαντικά υποεκπαιδευμένο, οπότε θεωρήθηκε πως μπορεί να φτάσει ή να ξεπεράσει τις επιδόσεις κάθε μοντέλου που δημοσιεύθηκε μετά από αυτό [36] [10].
- deepset/minilm-uncased-squad2 [11]: Το μοντέλο MiniLM ένα μοντέλο μετασχηματισμών προτάσεων: Χαρτογραφεί προτάσεις και παραγράφους σε έναν πυκνό διανυσματικό χώρο 384 διαστάσεων και μπορεί να χρησιμοποιηθεί για εργασίες όπως η ομαδοποίηση ή η σημασιολογική αναζήτηση. Το έργο αποσκοπεί στην εκπαίδευση μοντέλων ενσωμάτωσης προτάσεων σε πολύ μεγάλα σύνολα δεδομένων σε επίπεδο προτάσεων, χρησιμοποιώντας έναν αυτοεπιβλεπόμενο στόχο αντιθετικής μάθησης. Χρησιμοποιεί το προεκπαιδευμένο μοντέλο preimers/MiniLM-L6-H384-uncased και έγινε fine-tuned σε ένα σύνολο δεδομένων 1B ζευγών προτάσεων. Χρησιμοποιεί έναν στόχο αντιθετικής μάθησης: δεδομένης μιας πρότασης από το ζεύγος, το μοντέλο θα πρέπει να προβλέψει ποια από ένα σύνολο τυχαία δειγματοληπτικών άλλων προτάσεων, ήταν πραγματικά ζευγαρωμένη με αυτήν στο σύνολο δεδομένων. Το μοντέλο προορίζεται να χρησιμοποιηθεί ως κωδικοποιητής προτάσεων και σύντομων παραγράφων. Δεδομένου ενός κειμένου εισόδου, εξάγει ένα διάνυσμα που αποτυπώνει τις σημασιολογικές πληροφορίες. Το διάνυσμα πρότασης μπορεί να χρησιμοποιηθεί για εργασίες ανάκτησης πληροφοριών, ομαδοποίησης ή ομοιότητας προτάσεων [11].
- albert-xxlarge-v2 [12]: Το μοντέλο ALBERT (XXL) είναι ένα μεγάλο και ισχυρό state of the art μοντέλο. Είναι το πιο ακριβές μοντέλο open source ερωτοαπαντήσεων. Η υπολογιστική ισχύς που απαιτείται το καθιστά μη πρακτικό στις περισσότερες εφαρμογές [42]. Είναι ένα προεκπαιδευμένο μοντέλο στην αγγλική γλώσσα με χρήση ενός στόχου μασκοφόρου γλωσσικής μοντελοποίησης (MLM). Όλα τα μοντέλα ALBERT, είναι uncased: δεν κάνει διαφορά μεταξύ κεφαλαίων και πεζών γραμμάτων [12].
- xlm-roberta-base [13]: Το XLM RoBERTa είναι μια πολύγλωσση έκδοση του RoBERTa. Έχει προεκπαιδευτεί σε 2,5TB φιλτραρισμένων δεδομένων CommonCrawl που περιέχουν 100 γλώσσες. Πιο συγκεκριμένα, προ-εκπαιδεύτηκε με τον στόχο Masked language modeling (MLM). Λαμβάνοντας μια πρόταση, το μοντέλο καλύπτει τυχαία το 15% των λέξεων στην είσοδο και στη συνέχεια τρέχει ολόκληρη την καλυμμένη πρόταση μέσω του μοντέλου και πρέπει να προβλέψει τις καλυμμένες λέξεις. Αυτό διαφέρει από τα παραδοσιακά επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) που συνήθως βλέπουν τις λέξεις τη μία μετά την άλλη, ή από αυτοπαλινδρομικά μοντέλα όπως το GPT που καλύπτουν εσωτερικά τα μελλοντικά tokens. Επιτρέπει στο μοντέλο να μάθει μια αμφίδρομη αναπαράσταση της πρότασης. Με αυτόν τον τρόπο, το μοντέλο μαθαίνει μια εσωτερική αναπαράσταση 100 γλωσσών που μπορεί στη συνέχεια να χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών χρήσιμων για μεταγενέστερες εργασίες. Αξίζει να σημειωθεί ότι αυτό το μοντέλο στοχεύει κυρίως στη λεπτομερή προσαρμογή σε εργασίες που χρησιμοποιούν ολόκληρη την πρόταση (ενδεχομένως καλυμμένη) για τη λήψη αποφάσεων, όπως η ταξινόμηση ακολουθιών, η ταξινόμηση συμβόλων ή η απάντηση ερωτήσεων [43].
- microsoft/deberta-v3-base [14]: Το DeBERTa V3 αποτελεί βελτίωση του DeBERTa με χρήση προεκπαίδευσης τύπου ELECTRA με διαβαθμισμένη-διασταυρούμενη ανταλλαγή ενσωμάτωσης. Το DeBERTa βελτιώνει τα μοντέλα BERT και RoBERTa χρησιμοποιώντας αποδιαταγμένη προσοχή και βελτιωμένο αποκωδικοποιητή μάσκας. Με αυτές τις δύο βελτιώσεις, το DeBERTa υπερτερεί του RoBERTa στην πλειονότητα των εργασιών NLU με δεδομένα εκπαίδευσης 80GB [44].
- deepset/gelectra-base-germanquad [15]: Το Gelectra είναι ένα εκπαιδευμένο γερμανικό μοντέλο ερωτοαπαντήσεων [45] βασισμένο στο μοντέλο gelectra-base. Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι το GermanQuAD, ένα νέο, γερμανόφωνο σύνολο δεδομένων το οποίο αξιολογήθηκε και δημοσιεύθηκε στο διαδίκτυο. Το σύνολο δεδομένων εκπαίδευσης είναι μονόδρομα σημασμένο και περιλαμβάνει 11518 ερωτήσεις και 11518 απαντήσεις. Το σύνολο δεδομένων δοκιμής είναι τριαδικά σημασμένο, με 2204 ερωτήσεις και 6536 απαντήσεις μετά την αφαίρεση 76 λανθασμένων απαντήσεων [15].

Από τις προδιαγραφές των μοντέλων, το RoBERTa, συγκριτικά με τα άλλα δύο μοντέλα της Αγγλικής γλώσσας, φαίνεται να υπερτερεί ως προς το ότι είναι all-around και είναι βελτιστοποιημένο στην κάρτα γραφικών Nvidia V100. Ωστόσο, αναμένεται να υστερεί όμως σε ταχύτητα και ακρίβεια [42].

Το MiniLM, σύμφωνα με τα στοιχεία της deepset [16], υπερτερεί έναντι των άλλων δύο στο γεγονός ότι είναι 40% μικρότερο και 50% ταχύτερο, με καλύτερη ακρίβεια σε σχέση με το μοντέλο που βασίζεται στο BERT [42]. Υστερεί όμως σε σχέση με τα άλλα μοντέλα στην ακρίβεια.

Τα μοντέλα XLM RoBERTa και DeBERTa V3 είναι πολυγλωσσικά μοντέλα, που είναι εκπαιδευμένα σε πάνω από μια γλώσσες.

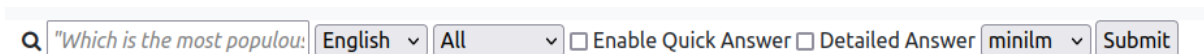
Επιλογή των μοντέλων

Για την επιλογή των μοντέλων ελήφθησαν υπόψη το μέγεθος, η εξειδίκευση και η γλώσσα πάνω στην οποία έχουν εκπαιδευτεί. Τα 3 μοντέλα που είναι πάνω στην Αγγλική γλώσσα, διαφέρουν ως προς το μέγεθος. Αυτό έχει επίπτωση τόσο στην ποιότητα των απαντήσεων του όσο και στην ταχύτητα εκτέλεσης του προγράμματος. Αναμένεται το μοντέλο ALBERT να είναι αρκετά αργό, τόσο που να μην είναι αποδεκτό για την υλοποίηση αυτή. Ωστόσο, τα αποτελέσματά του θα πρέπει να είναι πολύ ακριβή. Το μοντέλο MiniLM από την άλλη είναι ιδιαίτερα μικρό, οπότε οι απαντήσεις επρόκειτο να είναι χειρότερης ποιότητας, αλλά να επιστρέφονται σε πολύ μικρό χρονικό διάστημα. Το μοντέλο RoBERTa υπολογίζεται ότι θα βρίσκεται σε κάποια ενδιάμεση κατάσταση με επιδόσεις, ως προς την ποιότητα και τον χρόνο, μεταξύ των δύο προηγούμενων. Τα δυο άλλα μοντέλα είναι πολυγλωσσικά. Επιλέχθηκαν για να δώσουν ποικιλία στις απαντήσεις και να αποκτήσει η σύγκριση κάποια διαφορετική οπτική.

4.4 Διεπαφή Χρήστη

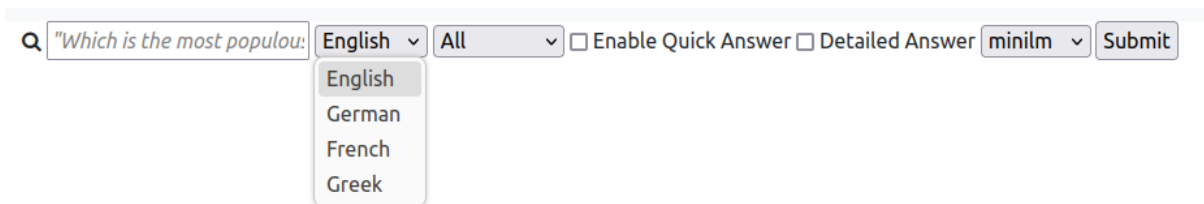
Η διεπαφή του χρήστη χωρίζεται σε δύο τμήματα. Το άνω τμήμα είναι ορατό στον χρήστη ανά πάσα χρονική στιγμή, ενώ το κάτω τμήμα είναι ορατό μόνο κατά την εμφάνιση των αποτελεσμάτων.

Το άνω τμήμα (Σχήμα 4.8) αποτελείται από την μπάρα αναζήτησης όπου ο χρήστης εισάγει το ερώτημά του. Ακριβώς στα δεξιά είναι ορατές οι επιλογές που σχετίζονται με το ερώτημα. Αυτές οι επιλογές έχουν αναλυθεί στο κεφάλαιο 4.2.1, τα Ορίσματα Χρήστη. Είναι τα question, language (Σχήμα 4.9), website (Σχήμα 4.10), model (Σχήμα 4.11), enable quick answer και detailed answer. Ακριβώς δίπλα βρίσκεται το κουμπί submit με το οποίο ο χρήστης μπορεί να υποβάλει το ερώτημά του.



Q "Which is the most populous: English All ☐ Enable Quick Answer ☐ Detailed Answer minilm Submit

Σχήμα 4.8: Η μπάρα που δέχεται το ερώτημα του χρήστη και οι επιλογές



Q "Which is the most populous: English All ☐ Enable Quick Answer ☐ Detailed Answer minilm Submit

- English
- German
- French
- Greek

Σχήμα 4.9: Η μπάρα με το μενού επιλογής γλώσσας

Q "Which is the most populou: English All Enable Quick Answer Detailed Answer minilm Submit

All
Wikipedia
in.gr

Σχήμα 4.10: Η μπάρα με το μενού επιλογής συγκεκριμένης ιστοσελίδας ως πηγή

Q "Which is the most populou: English All Enable Quick Answer Detailed Answer minilm Submit

roberta
minilm
albert

Σχήμα 4.11: Η μπάρα με το μενού επιλογής μοντέλου

Το κάτω τμήμα αποτελείται από το τμήμα των αποτελεσμάτων. Σε περίπτωση που ο αλγόριθμος εκτελεί όλα τα βήματα, τότε ο χρήστης έχει πρόσβαση σε όλα τα στάδια των αποτελεσμάτων. Στα πρώτα βήματα τυπώνει ο αλγόριθμος την απάντηση μαζί με την πηγή της. Στο τελευταίο βήμα τυπώνει ο αλγόριθμος τις καλύτερες απαντήσεις μαζί με το σκορ της απάντησης του μοντέλου, το πλαίσιο στο οποίο βρέθηκε αυτή η απάντηση καθώς και τον υπερσύνδεσμο για την ιστοσελίδα στην οποία ανήκει η πηγή.

Στην αρχή εμφανίζεται η ιστοσελίδα που χρησιμοποιήθηκε ως πηγή στο πρώτο στάδιο του αλγορίθμου (Σχήμα 4.12).

FlaskApp SkapisBot Results About

Q Who is the President of the English All Enable Quick Answer Detailed Answer roberta Submit

Result 1:
https://en.wikipedia.org/wiki/President_of_the_United_States

Result 2:

The answer is: **Joe Biden** with as score of 0.8686485886573792
The answer is: **President Barack Obama** with as score of 0.4096468389034271
The answer is: **Barack Obama** with as score of 0.2043895274400711
The answer is: **be a resident in the United States for at least 14 years** with as score of 0.05172586441040039

Result 3:

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Donald Trump** with as score of 0.2932969331741333 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
context: aphic, not a presidential seal. The 2019 spoof seal On July 23, 2019, President Donald Trump, the forty-fifth president, gave an address to young Republicans at t

The answer is: **Donald Trump** with as score of 0.2316812127828598 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
context: a presidential seal. The 2019 spoof seal On July 23, 2019, President Donald Trumo, the forty-fifth president, oave an address to vouno Republicans at t

Σχήμα 4.12: Πρώτη πηγή

Στη συνέχεια εντοπίζονται οι απαντήσεις, με τα σκορ, σε αυτό το στάδιο (Σχήμα 4.13).

FlaskApp SkapisBot Results About

Q *Who is the President of the* English All ☐ Enable Quick Answer ☐ Detailed Answer roberta Submit

Result 1:
https://en.wikipedia.org/wiki/President_of_the_United_States

Result 2:

The answer is: **Joe Biden** with as score of 0.8686485886573792
 The answer is: **President Barack Obama** with as score of 0.4096468389034271
 The answer is: **Barack Obama** with as score of 0.2043895274400711
 The answer is: **be a resident in the United States for at least 14 years** with as score of 0.05172586441040039

Result 3:

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Donald Trump** with as score of 0.2932969331741333 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
 context: aphic, not a presidential seal. The 2019 spoof seal On July 23, 2019, President Donald Trump, the forty-fifth president, gave an address to young Republicans at t

The answer is: **Donald Trump** with as score of 0.2316812127828598 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
 context: a presidential seal. The 2019 spoof seal On Julv 23. 2019. President Donald Trumo. the forty-fifth president. oave an address to young Republicans at t

Σχήμα 4.13: Πρώτα αποτελέσματα

Τέλος, εμφανίζονται τα συνολικά αποτελέσματα. Η απάντηση σε έντονη γραφή, το σκορ, η ιστοσελίδα που χρησιμοποιήθηκε σαν πηγή, και το context της απάντησης. Οι απαντήσεις που προήλθαν από το πρώτο βήμα επιτηδες παρουσιάζονται διπλότυπες, για να φανεί ότι το σκορ είναι ίδιο σε κάθε βήμα (Σχήμα 4.14).

FlaskApp SkapisBot Results About

Q *Who is the President of the* English All ☐ Enable Quick Answer ☐ Detailed Answer roberta Submit

Result 1:
https://en.wikipedia.org/wiki/President_of_the_United_States

Result 2:

The answer is: **Joe Biden** with as score of 0.8686485886573792
 The answer is: **President Barack Obama** with as score of 0.4096468389034271
 The answer is: **Barack Obama** with as score of 0.2043895274400711
 The answer is: **be a resident in the United States for at least 14 years** with as score of 0.05172586441040039

Result 3:

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **Joe Biden** with as score of 0.8686485886573792 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: dividuals have served 46 presidencies spanning 58 four-year terms.[C] Joe Biden is the 46th and current president of the United States, having assume

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Barack Obama** with as score of 0.4096468389034271 in website: https://en.wikipedia.org/wiki/President_of_the_United_States
 context: Setting the agenda President Barack Obama delivers the 2015 State of the Union Address with Vice President Joe

The answer is: **President Donald Trump** with as score of 0.2932969331741333 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
 context: aphic, not a presidential seal. The 2019 spoof seal On July 23, 2019, President Donald Trump, the forty-fifth president, gave an address to young Republicans at t

The answer is: **Donald Trump** with as score of 0.2316812127828598 in website: https://en.wikipedia.org/wiki/Seal_of_the_president_of_the_United_States
 context: a presidential seal. The 2019 spoof seal On Julv 23. 2019. President Donald Trumo. the forty-fifth president. oave an address to young Republicans at t

Σχήμα 4.14: Συγκεντρωτικά αποτελέσματα

4.5 Ιδιωτικότητα και Προσωποποιημένο Σύστημα

Κατά την χρήση του συστήματος, κάθε φορά που ο χρήστης κάνει ένα ερώτημα και καλείται πληροφορία, αποθηκεύεται το σύνολο αυτών των δεδομένων στο Document Store. Οπότε, όσο αυξάνεται το πλήθος των ερωτήσεων αυξάνεται και το πλήθος της πληροφορίας που είναι διαθέσιμη πιο άμεσα στον χρήστη. Η βάση δεδομένων με κάθε ερώτημα χτίζεται πάνω στις πληροφορίες που ο κάθε χρήστης αναζητά και γίνεται όλο και πιο προσωποποιημένη προσφέροντας αμεσότητα, και ως προς την ταχύτητα απόκρισης αλλά και ως προς την προέλευση της πληροφορίας.

Ενώ τα δεδομένα στο Document Store, στην παρούσα εργασία χρησιμοποιήθηκε το Elasticsearch, μπορούν να υποστούν επεξεργασία και να βελτιωθούν για να αντληθούν ωφέλιμες πληροφορίες, αυτές οι μετρήσεις, μετρικές και πληροφορίες, πρέπει συχνά να μεταδίδονται σε άλλες πλατφόρμες ή συσκευές, έτσι ώστε ο χρήστης να μπορεί να επιτύχει και πιο ιδιωτικό αποτέλεσμα αλλά και πιο προσωποποιημένο.

Έτσι, η ενσωμάτωση και εξαγωγή δεδομένων από το Elasticsearch είναι ένα φαινόμενο που χρησιμοποιείται ευρέως από εταιρίες και οργανισμούς για να αποκτήσουν τα δεδομένα στις επιθυμητές θέσεις αποθήκευσης.

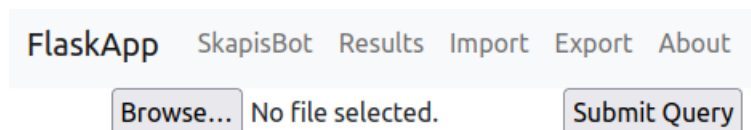
Η διαδικασία είναι πολύ απλή. Είτε με λογισμικό κάποιος τρίτης εταιρίας, είτε μέσω εξαγωγής του Elasticsearch Dump αρχείου, είτε και με απευθείας χειραγώγηση των δεδομένων μέσω κάποιος γλώσσας προγραμματισμού, είναι δυνατή η εξαγωγή ολόκληρης της βάσης δεδομένων που έχει χτιστεί προσωπικά για τον κάθε χρήστη. Το αρχείο που δίνεται σαν προϊόν αυτής της διαδικασίας είναι ένα JSON αρχείο είτε το Dump το ίδιο.

Ενώ οι παραπάνω μέθοδοι για χειροκίνητη εξαγωγή είναι εύκολοι να εφαρμοστούν και να ακολουθηθούν, υπάρχουν προφανείς περιορισμοί. Αρχικά, απαιτούνται τεχνικές γνώσεις για να πραγματοποιηθεί η εγκατάσταση απαιτούμενων προσθέτων και η εφαρμογή των απαραίτητων βημάτων. Παρόλο που πολλοί επαγγελματίες μπορεί να θεωρούν αυτό το έργο εφικτό, οι χρήστες που επιθυμούν να εξάγουν τα δεδομένα τους μπορεί να μην είναι κατάλληλα καταρτισμένοι για την εκτέλεση αυτής της διαδικασίας.

Δεύτερον, κάθε χειροκίνητη ενσωμάτωση ενέχει την πιθανότητα σφαλμάτων και ασυνεπούς ή μη έγκυρου κώδικα που μπορεί να προκαλέσει προβλήματα στη γενική διαδικασία. Εάν εφαρμοστεί εσφαλμένα, υπάρχει επίσης η πιθανότητα να χαθούν δεδομένα και να καταστραφούν σημαντικά αρχεία. Αυτά τα ζητήματα μπορούν να επιλυθούν με τη χρήση πλατφορμών που δεν βασίζονται στην εφαρμογή κώδικα, και μπορούν να αυτοματοποιήσουν τη διαδικασία εξαγωγής.

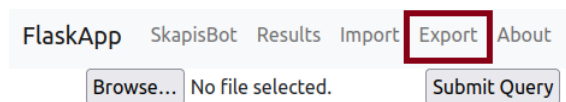
Για αυτόν το λόγο αναπτύχθηκε μια αυτόματη διαδικασία που ο χρήστης μπορεί να εξάγει ολόκληρο το Elasticsearch που έχει "χτίσει", και να το μεταφέρει στο ίδιο ή άλλο σύστημα. Η εισαγωγή εξωτερικού συστήματος είναι και πάλι δυνατή με μια αυτόματη διαδικασία. Ο χρήστης μπορεί να επισκεφτεί τις καρτέλες Import και Export αντίστοιχα, και να ανεβάσει ή να κατεβάσει το σύνολο των δεδομένων που αυτός θέλει.

Στην καρτέλα Import ο χρήστης έχει την επιλογή να ανεβάσει το αρχείο του είτε σε dump μορφή, είτε σε JSON αρχείο (Σχήμα 4.15). Έπειτα, μπορεί να πάει στην σελίδα αναζήτησης και να κάνει ερωτήματα πάνω στα δικά του δεδομένα.



Σχήμα 4.15: Import

Αντίστοιχα, πατώντας πάνω στην καρτέλα Export (Σχήμα 4.16) κατεβαίνουν αυτόματα οι πληροφορίες που έχουν συλλεχθεί από τα ερωτήματα του χρήστη.



Σχήμα 4.16: Export

Με αυτόν τον τρόπο επιτυγχάνονται και η ιδιωτικότητα και το προσωποποιημένο σύστημα. Ο χρήστης πλέον μπορεί να χρησιμοποιεί τη δικιά του βάση δεδομένων για την εξόρυξη της πληροφορίας. Δεν χρειάζεται να αφήνονται ψηφιακά αποτυπώματα με αναζητήσεις μέσω API, ή άλλες τεχνικές, ενώ την ίδια στιγμή η πληροφορία που είναι διαθέσιμη είναι η ακριβής εκείνη που επιθυμεί ο χρήστης.

4.6 Προσαρμογή σε ιστοσελίδα

Το σύστημα παρέχει τη δυνατότητα στον χρήστη να δεχτεί απάντηση με πληροφορία από μια συγκεκριμένη ιστοσελίδα. Όπως φαίνεται και στο Σχήμα 4.10, ο χρήστης μπορεί να επιλέξει ένα συγκεκριμένο website. Αυτή η λειτουργία προσομοιώνει τη μεταφορά του συστήματος, ως add-on, σε συγκεκριμένες ιστοσελίδες. Αντικαθιστώντας την παραδοσιακή αναζήτηση, που πολλές φορές τα αποτελέσματα είναι φτωχά, το σύστημα δύναται να αποκρίνεται με εξειδικευμένη πληροφορία, καλύπτοντας ταυτόχρονα την ανάγκη για σύντομη και ακριβή απάντηση, και την ανάγκη για το ευρύτερο πλαίσιο της πληροφορίας (άρθρο, λήμμα, φωτογραφία ή άλλο πολυμέσο).

Κεφάλαιο 5

Πειράματα

5.1 Σύγκριση μοντέλων

Με τον Reader FARMReader, όπως αναφέρθηκε και νωρίτερα, χρησιμοποιούνται 6 μοντέλα. Τα 3 πρώτα είναι εκπαιδευμένα πάνω στην Αγγλική γλώσσα ενώ τα 3 τελευταία είναι πολυγλωσσικά (multi-lingual) :

- RoBERTa
- MiniLM
- ALBERT (XXL)
- XLM RoBERTa
- DeBERTa V3
- Gelectra

5.1.1 Σύγκριση μοντέλων ως προς την επίδοση

Για τη σύγκριση αυτών των μοντέλων πραγματοποιήθηκαν τα ίδια ερωτήματα και χρησιμοποιήθηκαν οι ίδιες πηγές, οι οποίες βρίσκονται αυτόματα με τον ίδιο αλγόριθμο, έτσι ώστε τα αποτελέσματα των 3 μοντέλων να είναι συγκρίσιμα. Ως μέτρο σύγκρισης χρησιμοποιήθηκαν το σκορ της κάθε απάντησης, ο χρόνος εκτέλεσης ολόκληρου του αλγορίθμου αλλά και η υποκειμενική ορθότητα του αποτελέσματος. Το σκορ της κάθε απάντησης είναι μια μετρική που επιστρέφει το ίδιο το μοντέλο. Παίρνει τιμές από 0 έως 1, όπου το 1 είναι το άριστο. Ο χρόνος εκτέλεσης του αλγορίθμου υπολογίζεται και ως σύνολο αλλά και σε διαφορετικά στάδια. Τέλος, το κριτήριο υποκειμενικής ορθότητας είναι η υποκειμενική άποψη του γράφοντα σχετικά με την ορθότητα των απαντήσεων που επιστρέφει το κάθε μοντέλο. Οι απαντήσεις αξιολογήθηκαν ως προς το περιεχόμενό τους και κρίθηκαν ορθές ή λανθασμένες ανάλογα αν ταυτίζονται με τη αναμενόμενη απάντηση. Σε ορισμένες περιπτώσεις που η απάντηση στο ερώτημα δεν είναι αντικειμενική, θεωρήθηκαν ως ορθές όλες οι απαντήσεις που θα μπορούσαν να είναι αποδεκτές υπό οποιοδήποτε πλαίσιο.

5.1.2 Ερωτήματα

Για τη σύγκριση των μοντέλων πραγματοποιήθηκαν 49 ερωτήματα. Τα μοντέλα xlm-roberta-base και microsoft/deberta-v3-base είχαν απογοητευτικά αποτελέσματα, οπότε και απορρίφθηκαν γρήγορα. Θεωρήθηκε πως δεν αξίζει να γίνεται σύγκριση με τα άλλα μοντέλα, καθώς τα αποτελέσματα είχαν πολύ χαμηλή ποιότητα (χαμηλά σκορ και απαντήσεις που δεν ικανοποιούν). Το μοντέλο ahotrod/albert_xxlargev1_squad2_512 έδειξε εξαιρετικές επιδόσεις, με απαντήσεις με πολύ υψηλά σκορ, ωστόσο οι χρόνοι εκτέλεσης ήταν εξαντλητικοί που έκαναν αδύνατη την εκτέλεσή του πολλές φορές. Για αυτόν τον λόγο με το μοντέλο ahotrod/albert_xxlargev1_squad2_512 έγιναν 7 ερωτήματα, και στα άλλα δυο πολυγλωσσικά έγιναν μόνο 3 ερωτήματα στα Αγγλικά και 3 στα Γερμανικά. Στα μοντέλα deepset/minilm-uncased-squad2, deepset/roberta-

base-squad2 και deepset/gelectra-base-germanquad έγιναν και τα 49 ερωτήματα που χρησιμοποιήθηκαν.

Η επιλογή των ερωτημάτων έγινε με τέτοιο τρόπο ώστε να καλυφθούν όσο το δυνατόν περισσότερα είδη ερωτήσεων. Επιλέχθηκαν ερωτήματα κλειστού και ανοιχτού τύπου, ερωτήματα μονολεκτικών απαντήσεων, ερωτήματα υποκειμενικά ή αντικειμενικά, ερωτήματα μαθηματικών. Επιλέχθηκαν και 3 ερωτήματα στα Γερμανικά, για να συγκριθούν και τα πολυγλωσσικά μοντέλα. Η προσπάθεια του γράφοντα να καλύψει διαφορετικού τύπου ερωτημάτων οδήγησε στα 49 ακόλουθα:

1. Welche ist die Hauptstadt von Deutschland?
2. Welche Geschwindigkeit hat das Licht?
3. Wann fand die Mondlandung statt?
4. Who is Cristiano Ronaldo?
5. Which is the capital of Greece?
6. Who is the President of the USA?
7. How can I reset my password?
8. Provide me with a brief history of World War II.
9. Explain the concept of artificial intelligence.
10. What's the current exchange rate between USD and Euro?
11. Calculate 15 multiplied by 32.
12. What is the meaning of the word 'serendipity'?
13. Translate "Hello, how are you?" into French.
14. Tell me a fun fact about the planet Mars.
15. Explain the concept of blockchain technology.
16. How do I file my taxes online?
17. Can you provide recommendations for a good science fiction book to read?
18. What's the latest news in technology?
19. How can I improve my productivity at work?
20. Provide me with a summary of the movie "Inception."
21. What's the tallest mountain in the world?
22. Can you recommend a good workout routine for beginners?
23. Tell me a famous quote from Albert Einstein.
24. Provide a definition of the term "climate change."
25. How does photosynthesis work?
26. What are the basic principles of supply and demand in economics?
27. Can you explain the concept of quantum computing?
28. Who won the last World Series in baseball?
29. How do I troubleshoot a slow internet connection?
30. How can I prepare a perfect cup of coffee at home?
31. Can you recommend some good movies to watch this weekend?

32. What are the symptoms of COVID-19?
33. Explain the theory of relativity by Albert Einstein.
34. What's the process for applying for a student visa to the United States?
35. Who are the major characters in the movie series "Star Wars"?
36. What is the recipe for a classic margarita cocktail?
37. What are some tips for taking great photographs with a smartphone?
38. Where is Newtown Township?
39. What is The Dashuigou Formation?
40. Who is Evangelia Ritzaleou?
41. Where was the first Grand Prix victory of Jenson Button?
42. When did the first extension of the Athens Tram take place?
43. When did Socrates die?
44. Which is the most populous country.
45. How much is 1+1?
46. Which is the meaning of life?
47. Describe the earth's shape.
48. Who is Donald Trump?
49. Who is the inventor of the airplane?

5.1.3 Ποιότητα των Απαντήσεων

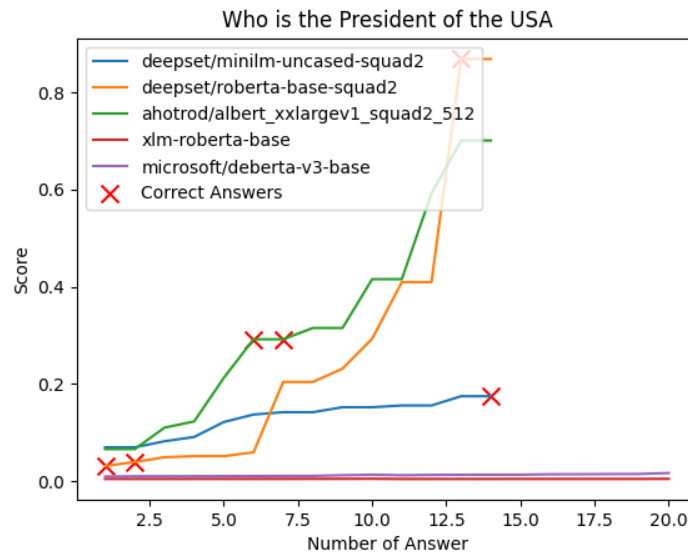
Αρχικά, θα πραγματοποιηθεί μια σύγκριση των 5 από τα 6 μοντέλα σε ένα κοινό ερώτημα, το "Who is the President of the USA". Στον Πίνακα 5.1 γίνεται μια σύγκριση των απαντήσεων που δόθηκαν. Τα δύο πολυγλωσσικά μοντέλα δεν έχουν καμία σωστή απάντηση, ενώ ταυτόχρονα το σκορ των απαντήσεων είναι εξαιρετικά χαμηλό. Αντιθέτως, τα δύο μοντέλα της Deepset και το μοντέλο albert έχουν δώσει πολύ υψηλότερα σκορ. Μάλιστα έχουν δοθεί και κάποιες ορθές απαντήσεις, με το ποσοστό ορθότητας να κυμαίνεται από 7,14% μέχρι 21,43%. Το σκορ του μοντέλου minilm είναι αρκετά χαμηλό αλλά και πάλι δίνει μια σωστή απάντηση.

Who is the President of the USA /Μοντέλο	Αριθμός απαντήσεων	Αριθμός ορθών απαντήσεων	Ποσοστό ορθών απαντήσεων(%)	Υψηλότερο Score απάντησης
deepset/minilm-uncased-squad2	14	1	7,14	0,175
deepset/roberta-base-squad2	14	3	21,43	0,869
ahotrod/albert_xxlargev1_squad2_512	14	2	14,29	0,701
xlm-roberta-base	20	0	0	0,005
microsoft/deberta-v3-base	20	0	0	0,017

Πίνακας 5.1: Μετρικές της ερώτησης "Who is the President of the USA"

Στο Σχήμα 5.1 παρουσιάζεται το σκορ των απαντήσεων με τις σωστές απαντήσεις. Στα αριστερά βρίσκονται οι απαντήσεις με το χαμηλότερο σκορ ενώ στα δεξιά αυτές με το υψηλότερο. Με το σημάδι "X" είναι σημειωμένες όλες οι απαντήσεις που έγιναν αποδεκτές. Όπως είναι ορατό, τα δυο μοντέλα που περιμέναμε τα καλύτερα αποτελέσματα, το albert και το roberta-base, έχουν πληθώρα απαντήσεων με υψηλά σκορ. Το παράδοξο σε αυτό το ερώτημα είναι πως το μοντέλο albert είχε τις μοναδικές του απαντήσεις σε απαντήσεις με χαμηλότερα σκορ. Το μοντέλο roberta-base είχε μια σωστή απάντηση στη δεύτερη θέση και 2 στις δύο τελευταίες. Το

μοντέλο είχε την μοναδική σωστή απάντηση στην πρώτη θέση. Από αυτό το ερώτημα συμπεραίνεται πως οι σωστές απαντήσεις μπορεί να κρύβονται σε αποφάνσεις του συστήματος με χαμηλότερη βαθμολογία, μερικές φορές και εξαιρετικά χαμηλή. Το ερώτημα, αν και απλοϊκό, κρύβει τον κίνδυνο πως στο διαδίκτυο κρύβονται πληθώρα λανθασμένων απαντήσεων. Εξαιτίας της συνεχούς εναλλαγής στο αξίωμα του προέδρου, ελλοχεύει ο κίνδυνος να εντοπίσει το σύστημα πληροφορίες παλαιότερες ή συμφραζόμενα που να μην είναι ξεκάθαρα στο εκάστοτε μοντέλο. Πολλές από τις εσφαλμένες απαντήσεις, η συντριπτική πλειοψηφία τους, ήταν ονόματα παλαιότερων προέδρων, με κυρίαρχα αυτά των τελευταίων.



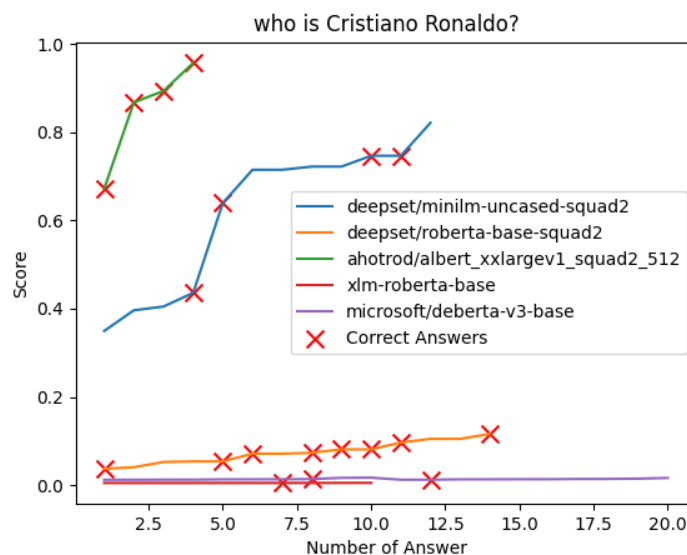
Σχήμα 5.1: Σύγκριση απαντήσεων στο ερώτημα "Who is the President of the USA"

Για την καλύτερη σύγκριση και κατανόηση της λειτουργίας των μοντέλων, πραγματοποιήθηκε μια δεύτερη ερώτηση ίδιου χαρακτήρα, η "who is Cristiano Ronaldo?". Σε αυτήν την περίπτωση το ερώτημα δεν μπορεί να παρερμηνευθεί, διότι η ταυτότητα του ατόμου που αναζητείται δεν μπορεί να έχει διυικό ή πολλαπλό χαρακτήρα. Τα αποτελέσματα είναι ιδιαίτερα ενδιαφέροντα σε κάθε πτυχή τους. Αρχικά, το μοντέλο albert τερμάτισε τον αλγόριθμό του στο πρώτο στάδιο του αλγορίθμου, γιατί το σκορ της απάντησης ήταν εξαιρετικά υψηλό (πάνω από το όριο *limit* που για την εκτέλεση των τεστ αυτών ορίστηκε στο 0,9). Και οι 4 απαντήσεις που δόθηκαν ήταν επιτυχείς. Το μοντέλο roberta μπορεί να μην είχε την ίδια επιτυχία ως προς το σκορ, αλλά πέτυχε 50% σωστές απαντήσεις. Η κορυφαία απάντηση, αν και με σκορ 0,116, ήταν σωστή. Αυτό το παράδειγμα κάνει προφανές ότι το σκορ δεν είναι πάντα ανάλογο της ορθότητας της απάντησης. Το μοντέλο minilm, σε αντίθεση με το πρώτο παράδειγμα που χρησιμοποιήθηκε, είχε αρκετά υψηλά σκορ, υψηλότερα του roberta, πετυχαίνοντας και 4 σωστές απαντήσεις, με πρώτες δύο από αυτές τη δεύτερη και την τρίτη απάντησή του. Τέλος, τα πολυγλωσσικά μοντέλα, παρά τα χαμηλά τους σκορ, επέστρεψαν 3 σωστές απαντήσεις συνολικά. Στον Πίνακα 5.2 είναι ορατά όλα τα στατιστικά του κάθε ξεχωριστού μοντέλου.

who is Cristiano Ronaldo? /Μοντέλο	Αριθμός απαντήσεων	Αριθμός ορθών απαντήσεων	Ποσοστό ορθών απαντήσεων(%)	Υψηλότερο Score απάντησης
deepset/minilm-uncased-squad2	14	4	28,57	0,821
deepset/roberta-base-squad2	14	8	57,14	0,116
ahotrod/albert_xxlargev1_squad2_512	4	4	100	0,956
xlm-roberta-base	20	1	5	0,005
microsoft/deberta-v3-base	20	2	10	0,016

Πίνακας 5.2: Μετρικές της ερώτησης "who is Cristiano Ronaldo?"

Στο Σχήμα 5.2 φαίνεται και η ακριβής θέση όλων των σωστών απαντήσεων σε σχέση με τα σκορ της κάθε απάντησης. Οι σωστές απαντήσεις σε αυτήν την ερώτηση ποίκιλαν. Κάποιες από αυτές επικεντρώνονταν στο πλήρες όνομα και την καταγωγή του ανθρώπου, άλλες στο επάγγελμά του, ποδοσφαιριστής, άλλες στο ποδοσφαιρικό σωματείο που αγωνίζεται, και άλλες στην κατάσταση του στην προσωπική του ζωή. Αυτό το ερώτημα οδήγησε στο συμπέρασμα πως το χαμηλό σκορ δεν εγγυάται και κακές ή εσφαλμένες απαντήσεις.



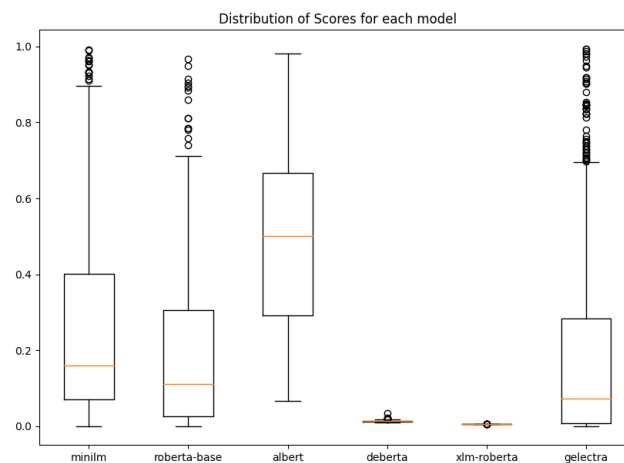
Σχήμα 5.2: Σύγκριση απαντήσεων στο ερώτημα "who is Cristiano Ronaldo?"

Τέλος, θα παρουσιαστεί και ένα ερώτημα όπου και τα 3 αγγλικά μοντέλα είχαν επιτυχία. Στο ερώτημα "Which is the capital of Greece", και τα 3 μοντέλα, minilm, roberta και albert, είχαν 4 μόνο απαντήσεις με πολύ υψηλές βαθμολογίες. Τα δύο πολυγλωσσικά μοντέλα και πάλι είχαν αξιολογήσεις κάτω από το 0,02. Οι μόνες 3 σωστές από τις 40 απαντήσεις που δόθηκαν ήρθαν από το μοντέλο microsoft/deberta-v3-base (Πίνακας 5.3).

Which is the capital of Greece /Μοντέλο	Αριθμός απαντήσεων	Αριθμός ορθών απαντήσεων	Ποσοστό ορθών απαντήσεων(%)	Υψηλότερο Score απάντησης
deepset/minilm-uncased-squad2	4	4	100	0,949
deepset/roberta-base-squad2	4	4	100	0,779
ahotrod/albert_xxlargev1_squad2_512	4	4	100	0,982
xlm-roberta-base	20	0	0	0,005
microsoft/deberta-v3-base	20	3	15	0,019

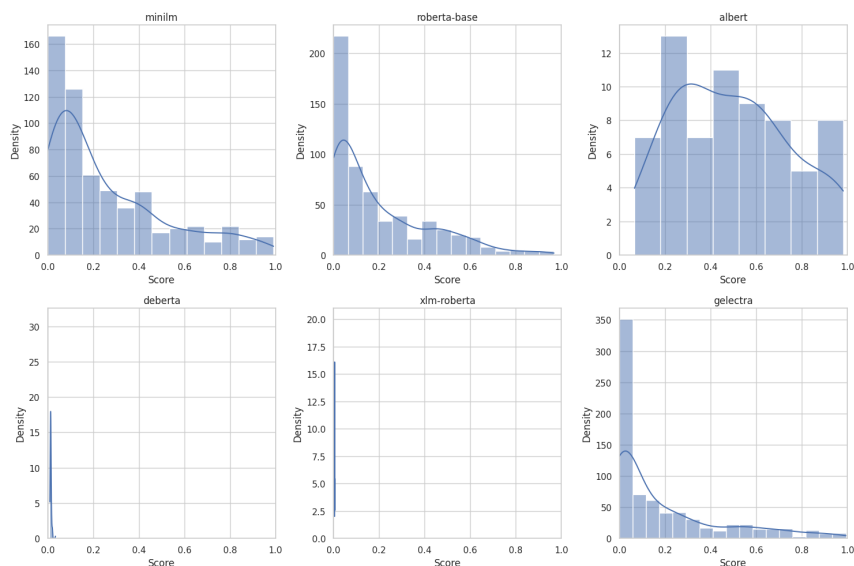
Πίνακας 5.3: Μετρικές της ερώτησης "Which is the capital of Greece"

Στη συνέχεια, θα παρουσιαστούν κάποιες μετρικές ως προς την ποιότητα και ποσότητα των απαντήσεων και στα 6 μοντέλα, αλλά και θα γίνει σύγκριση των αξιολογήσεων στο κάθε μοντέλο. Αρχικά, στα Σχήματα 5.3 και 5.4 συγκρίνονται και τα 6 μοντέλα ως προς της κατανομή των σκορ. Ως πρώτη παρατήρηση, τα δύο πολυγλωσσικά μοντέλα έχουν απογοητευτικές επιδόσεις. Το τρίτο πολυγλωσσικό μοντέλο, το gelectra, έχει επιδόσεις συγκρίσιμες με των τριών μοντέλων εκπαιδευμένων στην αγγλική γλώσσα. Το μοντέλο albert παρουσιάζει εξαιρετικές βαθμολογίες δημιουργώντας ένα τεράστιο χάσμα επίδοσης. Το μοντέλο minilm φαίνεται, με γυμνό μάτι, πως υπερτερεί του roberta έναντι στα συμπεράσματα των τριών ερωτημάτων που αναλύθηκαν παραπάνω.



Σχήμα 5.3: Κατανομή των σκορ του κάθε μοντέλου με τη χρήση Boxplot

Το σχήμα 5.4 δίνει μια καλύτερη εικόνα για το μοντέλο gelectra. Πλέον, είναι ορατό πως το μοντέλο έχει δώσει ένα τεράστιο ποσό κακών απαντήσεων (περίπου 350). Και πάλι γίνεται προφανές πως το minilm υπερτερεί ως έναν βαθμό του roberta, δίνοντας αυξημένες πολύ καλές απαντήσεις. Νωρίτερα μπορούσε να γίνει η ίδια παρατήρηση αν συγκρίνονταν τα άνω αυτιά των δύο Boxplots, με πιο έντονη τη συγκέντρωση απαντήσεων στην περιοχή με σκορ κοντά στην μονάδα.



Σχήμα 5.4: Κατανομή των σκορ του κάθε μοντέλου με τη χρήση ραβδογράμματος

Στον Πίνακα 5.4 γίνεται σύγκριση βασικών μετρικών. Ο μέσος όρος και η διάμεσος επιβεβαιώνουν τα σχόλια που έχουν γίνει σχετικά με την κατάταξη των μοντέλων. Τα μοντέλα minilm και roberta συνεχίζουν να βρίσκονται εξαιρετικά κοντά. Η τυπική απόκλιση είναι σχεδόν κοινή και για τα τέσσερα μοντέλα που επιστρέφουν κάποια αποδεκτά αποτελέσματα. Αυτό είναι και το αναμενόμενο, γιατί κάθε μοντέλο, ανεξάρτητα από την ποιότητα της μέσης απάντησής του, έχει και απαντήσεις κοντά στη μονάδα και στο μηδέν. Η λοξότητα είναι μια μετρική που υπολογίζει και αξιολογεί την ασυμμετρία μια κατανομής. Η τιμή 0 δηλώνει κανονική κατανομή, τιμές άνω του μηδενός δηλώνουν μεγαλύτερο βάρος στην αριστερή ουρά της κατανομής και αρνητικές τιμές δηλώνουν μεγαλύτερο βάρος στη δεξιά ουρά. Η κυρτότητα μιας κατανομής υπολογίζει το πόσο πλατιά ή στενή είναι η κατανομή συγκρινόμενη με την κανονική κατανομή. Όταν η τιμή της είναι

3, τότε είναι σαν την κανονική κατανομή. Όταν είναι μικρότερη του 3, τότε λέγεται πλατύκυρτη. Όταν είναι μεγαλύτερη του 3, λέγεται λεπτόκυρτη. Χωρίς να γίνεται αξιολόγηση των δυο πολυγλωσσικών μοντέλων με τα περιορισμένα αποτελέσματα, όλα τα υπόλοιπα μοντέλα έχουν θετική λοξότητα άρα και μεγάλη ασυμμετρία προς τα αριστερά. Το albert είναι σχετικά κοντά στο 0, αλλά τα υπόλοιπα τρία βρίσκονται μεταξύ 1 και 1,5. Αξίζει να παρατηρηθεί πως η ημι-κανονική κατανομή έχει λοξότητα 1. Επιπρόσθετα, όλα τα μοντέλα έχουν κυρτότητα μικρότερη του 3, καθώς το αριστερό άκρο τους είναι πιο έντονο. Το albert έχει τιμή -0,921, που το φέρνει πολύ κοντά στην ημικυκλική κατανομή Wigner που έχει τιμή -1. Τα υπόλοιπα μοντέλα έχουν τιμές από 0,29 μέχρι 1,1, μια περιοχή όπου βρίσκεται η λογιστική κατανομή με τιμή 1,2.

Μοντέλο	Average Score	Median Score	Variance	σ	Skewness	Kurtosis
minilm	0,263	0,161	0,066	0,256	1,122	0,294
roberta-base	0,2	0,111	0,046	0,215	1,250	0,855
albert	0,498	0,501	0,067	0,259	0,294	-0,921
xlm-roberta	0,0005	0,0005	0	0,003	3,167	16,699
deberta	0,014	0,013	0	0,0001	0,437	-0,342
gelectra	0,192	0,072	0,06	0,248	1,445	1,134

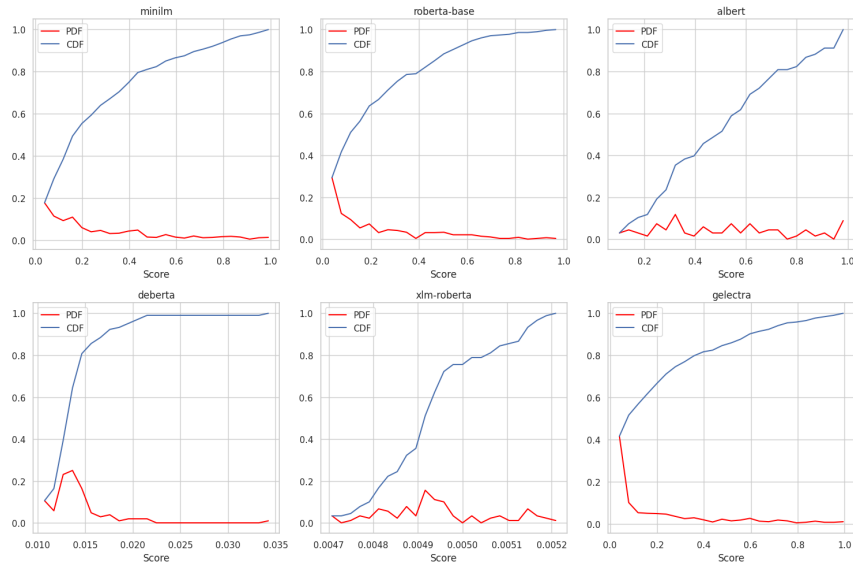
Πίνακας 5.4: Γενικές Μετρικές

Στον Πίνακα 5.5 παρουσιάζονται τα εκατοστημόρια των σκορ του κάθε μοντέλου. Στο 25ο εκατοστημόριο μόνο το μοντέλο albert έχει αξιοπρεπή αξιολόγηση, ενώ ταυτόχρονα έχει υψηλότερο σκορ στο 75ο. Για άλλη μια φορά το gelectra υστερεί έναντι των άλλων δύο μοντέλων, τα οποία παρουσιάζουν συγκρίσιμα αποτελέσματα, με το minilm να παίρνει τη μερίδα του λέοντος. Το διάστημα μεταξύ των δυο αυτών εκατοστημορίων είναι σχετικά παρόμοιο, αλλά δεν μας δίνει και πολλές πληροφορίες.

Μοντέλο	25th percentile	75th percentile	Interquartile range
minilm	0,071	0,401	0,33
roberta-base	0,026	0,305	0,279
albert	0,292	0,667	0,375
xlm-roberta	0,012	0,014	0,002
deberta	0,004	0,004	0,0001
gelectra	0,007	0,282	0,275

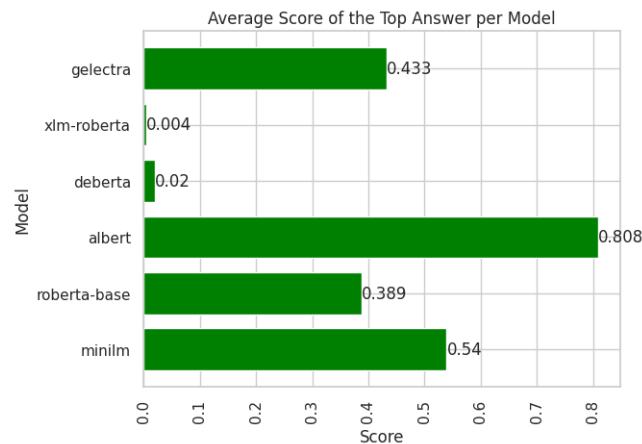
Πίνακας 5.5: Εκατοστημόρια

Στο Σχήμα 5.5 συγκρίνονται οι Συναρτήσεις Πιθανότητας και Αθροιστικής Πιθανότητας των κατανομών των σκορ. Δεν προκύπτουν νέα συμπεράσματα, αλλά και εδώ είναι προφανής η επικράτηση του μοντέλου albert, του οποίου η συνάρτηση αθροιστικής πιθανότητας είναι πολύ πιο ομαλή από τις υπόλοιπες. Η συνάρτηση του μοντέλου gelectra, παρ' ότι έχει συγκρίσιμη κλίση με τα υπόλοιπα μοντέλα, ξεκινάει από πολύ ψηλό σημείο, έχοντας παρά πολλές απαντήσεις με χαμηλό σκορ.



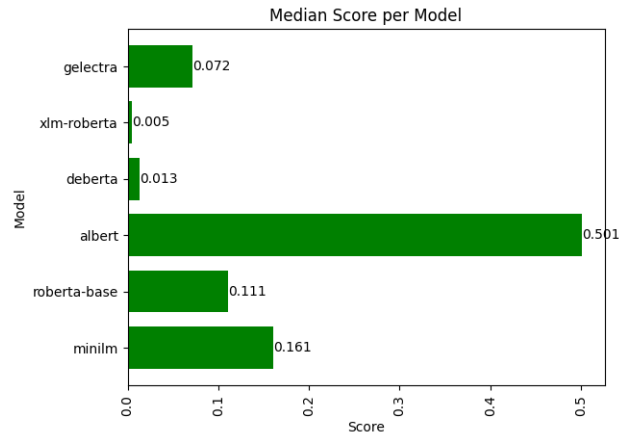
Σχήμα 5.5: Συνάρτηση Πιθανότητας και Αθροιστικής Πιθανότητας των σκορ

Στη συνέχεια, γίνεται προσπάθεια να συγκριθούν τα σκορ των κορυφαίων απαντήσεων. Ως πρώτο βήμα, επιλέγονται οι ερωτήσεις όπου το κάθε μοντέλο έχει τη χειρότερη σε σκορ κορυφαία απάντηση. Όλα τα μοντέλα εκτός του albert και του deberta δεν επέστρεψαν καν απάντηση. Η χειρότερη κορυφαία απάντηση του deberta ήταν 0,015 ενώ του albert 0,59. Αξίζει να σημειωθεί ότι το δείγμα για αυτά τα μοντέλα ήταν μικρότερο. Στη συνέχεια, στο Σχήμα 5.6 παρατίθενται οι μέσες τιμές των κορυφαίων απαντήσεων του κάθε μοντέλου. Η μέχρι τώρα αξιολόγηση επιβεβαιώνεται για μια ακόμα φορά, με τα 3 μοντέλα, gelectra, roberta και minilm να βρίσκονται και πάλι σχετικά κοντά.



Σχήμα 5.6: Μέση Κορυφαία Απάντηση

Στο Σχήμα 5.7 συγκρίνονται οι ενδιάμεσες τιμές των μοντέλων, με το albert να έχει συντριπτικό πλεονέκτημα.



Σχήμα 5.7: Ενδιάμεση Κορυφαία Απάντηση

Το προφανές ερώτημα που προκύπτει σε αυτό το σημείο είναι σε πόσες ερωτήσεις έχει την απάντηση με το μεγαλύτερο σκορ το κάθε μοντέλο. Στον Πίνακα 5.6 φαίνονται τα αθροιστικά στατιστικά. Το albert είχε πάνω από 42% επιτυχία στον περιορισμένο αριθμό των ερωτημάτων που χρησιμοποιήθηκε, καθιστώντας το το πιο πετυχημένο. Ανάμεσα στα άλλα ακολουθεί το minilm, και σε αντίθεση με τα στοιχεία των προηγούμενων διαγραμμάτων, το gelectra υπερτερεί του roberta-base. Αξίζει να σημειωθεί ότι σε ερωτήματα που υπήρξε ισοθαμία, ως κορυφαία απάντηση θεωρήθηκαν όλα τα ισοβαθμούντα μοντέλα.

Μοντέλο	Top Answer	Total Questions	Percentage of Success (%)
minilm	23	49	46,939
roberta-base	5	49	10,204
albert	3	7	42,857
xlm-roberta	0	6	0
deberta	0	6	0
gelectra	18	49	36,735

Πίνακας 5.6: Κορυφαία απάντηση ανά μοντέλο

Στο ίδιο πλαίσιο, είναι άξιο μελέτης το ποσοστό των σωστών απαντήσεων επί του συνόλου των απαντήσεων. Ο γράφων επιδίωξε μια πιο υποκειμενική σύγκριση των μοντέλων. Τα ερωτήματα, αν και διαφορετικής φύσεως το καθένα, δέχονται κάποια αντικειμενικότητα στις απαντήσεις τους. Ο γράφων αξιολόγησε όλες τις απαντήσεις των ερωτημάτων ως ορθές ή λανθασμένες. Στον Πίνακα 5.7 φαίνεται και πάλι ότι υπερτερεί το albert αλλά πλέον όχι με μεγάλη διαφορά. Το roberta-base, παρά τα χαμηλά σκορ και την αποτυχία του στις προηγούμενες αντικειμενικές μετρικές, έδωσε ένα πολύ μεγάλο ποσοστό (και πλήθος) σωστών απαντήσεων σε πάρα πολλά ερωτήματα. Παρ' ότι είχε χαμηλότερα σκορ από το minilm, κατάφερε να δώσει περίπου 46% σωστές απαντήσεις έναντι των 36% του minilm. Σε απόλυτα νούμερα το gelectra βρίσκεται πολύ κοντά στο minilm, με μόνο 8 απαντήσεις να τα χωρίζουν σε ένα δείγμα 600 με 750 απαντήσεων. Τα υπόλοιπα δύο πολυγλωσσικά μοντέλα είχαν σωστές απαντήσεις κάτω του 7%.

Μοντέλο	Correct Answer	Total Answers	Percentage of Success (%)
minilm	217	603	35,987
roberta-base	268	579	46,287
albert	33	68	48,529
xlm-roberta	7	107	6,542
deberta	5	90	5,556
gelectra	209	748	27,941

Πίνακας 5.7: Κορυφαία απάντηση ανά μοντέλο

Από τα παραπάνω κριτήρια και τις μετρικές, μπορούμε με σχετική ευκολία να εξάγουμε το συμπέρασμα πως το μοντέλο `ahotrod/albert_xxlargev1_squad2_512` είναι με διαφορά το καλύτερο ως προς και τα δύο κριτήρια, το σκορ και τις αντικειμενικές απαντήσεις, που χρησιμοποιούνται για την αξιολόγηση της ποιότητας της απάντησης. Στην πλειονότητα των περιπτώσεων έχει το μεγαλύτερο ποσοστό σωστών απαντήσεων αλλά και το μεγαλύτερο σκορ σε αυτές. Το μοντέλο `deepset/roberta-base-squad2` από την άλλη, έχει σκορ που το φέρνει στη τρίτη θέση, αλλά οι σωστές απαντήσεις του δίνουν μια άλλη εικόνα. Τα δεδομένα έδειξαν πως ανεξάρτητα από τα χαμηλά σκορ είναι ιδιαίτερα αξιόπιστο. Το μοντέλο `deepset/minilm-uncased-squad2` μπορεί να μην είναι δεύτερο στις υποκειμενικά σωστές απαντήσεις, αλλά τα σκορ του ήταν ιδιαίτερα ικανοποιητικά. Το `deepset/gelectra-base-germanquad`, συνδυάζοντας την ικανότητά του να απαντάει και Αγγλικά αλλά και ερωτήματα άλλων γλωσσών, βρίσκεται λίγο πιο πίσω, με τη "θυσία" στα σκορ αλλά και στην ορθότητα να είναι ένα σχετικά αποδεκτό trade-off. Αντιθέτως, τα μοντέλα `xlm-roberta-base` και `microsoft/deberta-v3-base`, τα άλλα δύο πολυγλωσσικά μοντέλα, είναι ιδιαίτερα απογοητευτικά στις επιδόσεις τους. Παρουσιάζουν πολύ χαμηλά σκορ και οι απαντήσεις τους είναι σωστές μόνο σε πολύ ορισμένα και τυχαία πλαίσια.

5.1.4 Σύγκριση μοντέλων ως προς την επίδοση (άλλες γλώσσες)

Για τη σύγκριση των μοντέλων έγιναν και κάποια ερωτήματα σε άλλες γλώσσες. Το ερώτημα "Welche ist die Hauptstadt von Deutschland?", που μεταφράζεται, "Ποια είναι η πρωτεύουσα της Γερμανίας" απαντήθηκε χρησιμοποιώντας τα μοντέλα `deepset/roberta-base-squad2`, `deepset/minilm-uncased-squad2`, `xlm-roberta-base` και `microsoft/deberta-v3-b`.

Με το μοντέλο `deepset/roberta-base-squad2` δόθηκαν έξι απαντήσεις, δύο από τις οποίες σωστές, ένα αποτέλεσμα που αποδείχθηκε αναπάντεχα πετυχημένο. Το μοντέλο `deepset/minilm-uncased-squad2` είχε 0 σωστές απαντήσεις στις 12 που έδωσε. Το μοντέλο `xlm-roberta-base` απέτυχε να δώσει κάποιο αποτέλεσμα, σωστό ή λανθασμένο. Με ακόμα χειρότερη επίδοση, το μοντέλο `microsoft/deberta-v3-b` επέστρεψε 8 απαντήσεις με μια μόνο, τη δεύτερη, σωστή. Το μοντέλο `deepset/gelectra-base-germanquad` έδωσε 20 απαντήσεις με 3 σωστές, τις πρώτες τρεις, με την πρώτη απάντηση να συγκεντρώνει σκορ 0,964, μια πολύ ικανοποιητική επίδοση.

Στη συνέχεια, πραγματοποιήθηκε το ερώτημα *Welche Geschwindigkeit hat das Licht?* που μεταφράζεται, "Ποια είναι η ταχύτητα του φωτός". Με το μοντέλο `deepset/roberta-base-squad2` δόθηκε πληθώρα απαντήσεων, με καμία από αυτές να είναι ακριβής. Ομοίως, το μοντέλο `deepset/minilm-uncased-squad2` είχε παρόμοια αποτελέσματα. Το μοντέλο `microsoft/deberta-v3-b` έδωσε μια σωστή απάντηση αλλά με αρκετά χαμηλό σκορ, ενώ το `xlm-roberta-base` έδωσε 2 σωστές απαντήσεις, τη 16η και τη 17η, με πολύ μικρό σκορ. Το μοντέλο `deepset/gelectra-base-germanquad` έδωσε 11 σωστές απαντήσεις, μια από τις καλύτερες επιδόσεις σε όλα τα ερωτήματα από όλα τα μοντέλα. Η κορυφαία απάντηση είχε σκορ 0,84.

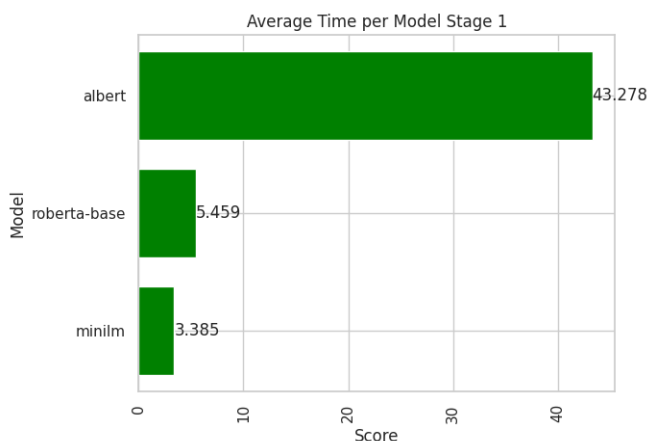
Τέλος, το ερώτημα *Wann fand die Mondlandung statt?*, που μεταφράζεται ως "Πότε πραγματοποιήθηκε η Προσελήνωση;" παρουσιάζει ενδιαφέρον. Τα μοντέλα `deepset/roberta-base-squad2` και `deepset/minilm-uncased-squad2` δεν έδωσαν καμία σωστή απάντηση, ομοίως και το `microsoft/deberta-v3-b`. Το `xlm-roberta-base` έδωσε 2 σωστές απαντήσεις, ενώ και πάλι το `deepset/gelectra-base-germanquad` έδωσε 8 σωστές απαντήσεις με χαμηλά σκορ αυτήν τη φορά, όλα κάτω του 0,15.

5.1.5 Σύγκριση μοντέλων ως προς τον χρόνο

Για τη σύγκριση των μοντέλων ως προς τον χρόνο μετρήθηκε το χρονικό διάστημα που περνάει για την εκτέλεση 2 τμημάτων του αλγορίθμου. Τα τμήματα αυτά είναι τα βήματα 11 με 18 και 20 με 23 χωρίς τη λειτουργία βάθους. Το αναμενόμενο αποτέλεσμα είναι ότι τα δύο στάδια θα έχουν ανάλογα αποτελέσματα, καθώς είναι ακριβώς η ίδια διαδικασία σε μεγαλύτερη κλίμακα. Σε γενικότερο πλαίσιο, το μοντέλο albert αναμένεται να είναι αρκετά πιο χρονοβόρο, ενώ τα μοντέλα roberta και minilm να έχουν συγκρίσιμους χρόνους εκτέλεσης, με το δεύτερο να είναι το πιο γρήγορο στο σύνολο των περιπτώσεων. Τα πολυγλωσσικά μοντέλα δεν συγκρίθηκαν χρονικά, καθώς παρουσιάστηκαν τεχνικά προβλήματα που αλλοίωναν τη σύγκριση.

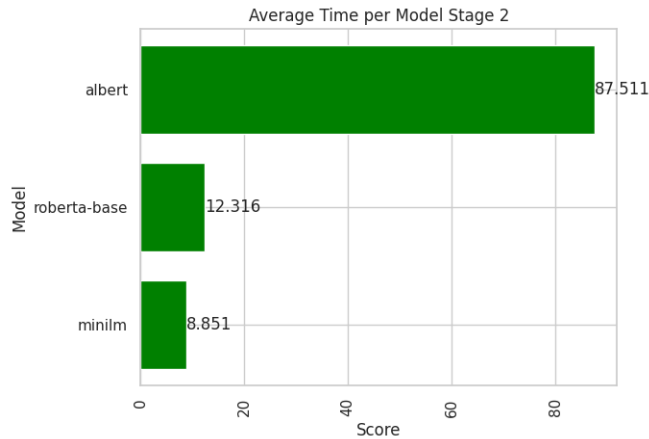
Στον Πίνακα 5.8 συγκρίνεται ο μέσος όρος διάρκειας εκτέλεσης του πρώτου σταδίου στο κάθε μοντέλο ενώ στο 5.9 ο για το δεύτερο στάδιο. Οι χρόνοι είναι σε δευτερόλεπτα και είναι αποτέλεσμα ολόκληρου του δείγματος εκτελέσεων.

Από τον Πίνακα 5.8 είναι προφανές το γεγονός πως το μοντέλο albert δεν είναι βιώσιμο. Το στάδιο 1 του αλγορίθμου είναι κατά μέσο όρο πάνω από 8 φορές πιο αργό από το roberta-base και πάνω από 13 φορές από το minilm. Ακόμα και αν υποθέσουμε πως το πρώτο στάδιο δίνει αρκετά ποιοτικές απαντήσεις για να παραλειφθεί το δεύτερο, το στάδιο 1 του albert είναι περίπου 4 φορές πιο αργό από το στάδιο 2 του roberta-base. Τα μοντέλα minilm και roberta-base έχουν μόλις 2 δευτερόλεπτα διαφορά στο πρώτο στάδιο αν και αυτό είναι περίπου 40% αύξηση χρόνου.



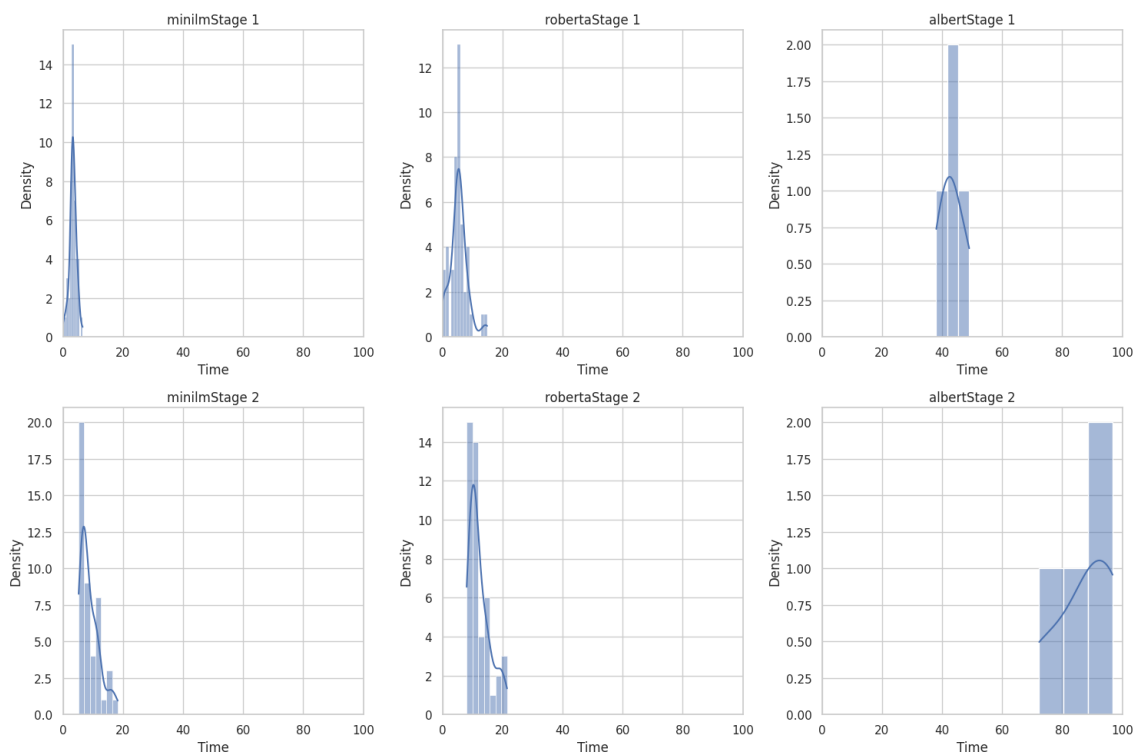
Σχήμα 5.8: Μέσοι χρόνοι εκτέλεσης κάθε μοντέλου στο 1ο στάδιο

Από τον Πίνακα 5.9 βλέπουμε πως το μοντέλο albert ξεφεύγει σε τρομακτικά μεγάλους χρόνους, παρ' ότι αναλογικά πλησιάζει τα άλλα δύο μοντέλα, καθώς πλέον είναι μόνο 7 και 10 πιο αργό. Τα μοντέλα minilm και roberta-base συνεχίζουν να έχουν μια διαφορά λίγο μεγαλύτερη από 3 δευτερόλεπτα. Οι συνολικοί χρόνοι απόκρισης του συστήματος δεν είναι αμελητέοι. 8,8 και 12 δευτερόλεπτα, για το minilm και το roberta-base αντίστοιχα, είναι αρκετός χρόνος για ένα ψηφιακό βοηθό. Οπότε θα πρέπει το σύστημα να βασίζεται σε μια απάντηση υψηλού σκορ από το πρώτο στάδιο.



Σχήμα 5.9: Μέσοι χρόνοι εκτέλεσης κάθε μοντέλου στο 2ο στάδιο

Παρατηρείται πως τα μοντέλα roberta και minilm έχουν σχετικά κοντινούς χρόνους στο πιο σύντομο βήμα. Αναμενόμενα με τις προβλέψεις που έγιναν νωρίτερα, η διαφορά των χρόνων μεγαλώνει αναλογικά στο επόμενο βήμα. Αντίστοιχα, το μοντέλο albert είναι εξαιρετικά αργό καθιστώντας το σχεδόν απαγορευτικό για χρήση σε μια εφαρμογή πραγματικού χρόνου όπως αυτή. Αξίζει να σημειωθεί πως πολλές φορές η κύρια χρονική καθυστέρηση προέρχεται από την ανάκληση των ιστοσελίδων. Για πιο αναλυτική κατανόηση των αποτελεσμάτων παρατίθενται παρακάτω κάποιες από τις μετρικές των συγκρίσεων που έγιναν. Στον Πίνακα 5.10 είναι ξανά ορατές οι παρατηρήσεις που έγιναν και νωρίτερα, με το τεράστιο χάσμα μεταξύ των 3 μοντέλων.



Σχήμα 5.10: Κατανομή των χρόνων του κάθε μοντέλου στο κάθε στάδιο με τη χρήση ραβδογράμματος

Στον Πίνακα 5.8 συγκρίνονται μετρικές των χρόνων για το πρώτο στάδιο και στον Πίνακα 5.9 για το δεύτερο.

Η μετρική της λοξότητας δηλώνει πως οι κατανομές είναι κοντά στην κανονική. Στο μοντέλο minilm έχει μεγαλύτερη δεξιά ουρά ενώ τα άλλα δύο μεγαλύτερη αριστερή. Αυτό σημαίνει πως στο μοντέλο minilm υπάρχει συγκέντρωση μικρότερων χρόνων ενώ στα υπόλοιπα μεγαλύτερων. Στο δεύτερο στάδιο, το albert έχει δεξιά ουρά και τα άλλα αριστερή. Η κυρτότητα είναι ξεκάθαρα πλατύκυρτη και στις 3 περιπτώσεις και στα 2 στάδια.

Στάδιο 1	Average Time	Median Time	Variance	σ	Skewness	Kurtosis
minilm	3,385	3,392	1,442	1,200	-0,207	1,031
roberta-base	5,459	5,458	8,818	2,969	0,816	2,045
albert	43,278	42,977	20,221	4,496	0,228	-1,003

Πίνακας 5.8: Πίνακας με μετρικές για τους χρόνους εκτέλεσης κάθε μοντέλου στο 1ο στάδιο

Στάδιο 2	Average Time	Median Time	Variance	σ	Skewness	Kurtosis
minilm	8,851	7,572	9,951	3,154	1,241	0,926
roberta-base	12,316	11,108	12,146	3,485	1,093	0,250
albert	87,511	90,444	120,105	10,959	-0,698	-1,087

Πίνακας 5.9: Πίνακας με μετρικές για τους χρόνους εκτέλεσης κάθε μοντέλου στο 2ο στάδιο

Στους πίνακες 5.10 και 5.11 παρατίθενται τα εκατοστημόρια, καθώς και οι ακραίοι χρόνοι. Σε όλες τις μετρικές είναι ορατή η αναλογία από το ένα στάδιο στο άλλο. Ωστόσο, το διάστημα μεταξύ 25ου και 75ου εκατοστημορίου στο μοντέλο albert δεν είναι τόσο μεγάλο όσο αναμενόταν. Οι ακραίες τιμές δείχνουν πως το στάδιο 2 φτάνει σε χρόνους που δεν θα είναι ευχάριστοι για έναν χρήστη. Ωστόσο, αυτές οι καθυστερήσεις μπορεί να οφείλονται σε αδύναμη σύνδεση στο διαδίκτυο κατά την εκτέλεση του πειράματος ως έναν βαθμό.

Stage 1	25th percentile	75th percentile	Interquartile range	Min Time	MaxTime
minilm	2,929	4,045	1,116	0,0879	6,591
roberta-base	4,463	6,234	1,770	0,1015	14,92
albert	41,504	44,750	3,245	38,096	49,058

Πίνακας 5.10: Πίνακας με εκατοστημόρια και ακραίες τιμές κάθε μοντέλου στο 1ο στάδιο

Stage 2	25th percentile	75th percentile	Interquartile range	Min Time	MaxTime
minilm	6,665	10,897	4,232	5,262	18,267
roberta-base	9,726	14,222	4,496	8,098	21,536
albert	83,149	94,805	11,655	72,359	96,793

Πίνακας 5.11: Πίνακας με εκατοστημόρια και ακραίες τιμές κάθε μοντέλου στο 2ο στάδιο

5.2 Σύγκριση Αξιολόγησης Ιστοσελίδων Βάθους

Για την αξιολόγηση των ιστοσελίδων βάθους έγινε χρήση 3 διαφορετικών τεχνικών, όπως ειπώθηκε και στο Κεφάλαιο 4. Στη συνέχεια, προστέθηκε και μια υποκειμενική μέθοδος αξιολόγησης του γράφοντα. Ο σκοπός αυτής της αξιολόγησης, χωρίς κάποια αντικειμενική συνάρτηση ή μετρική, είναι η μελέτη της ποιότητας των 3 τεχνικών που έχουν επιλεγεί. Για τη σύγκριση αυτή επιλέχθηκαν τυχαία 3 ερωτήματα. Τα αποτελέσματα είναι ορατά στους Πίνακες 5.12, 5.13 και 5.14.

Which is the capital of Greece	URL 1	URL 2	URL 3	URL 4
Pairwise Similarity	0,093825515	0,093825515	0,22012685	0,10173607
Existance	TRUE	TRUE	TRUE	TRUE
Meta Existance	TRUE	TRUE	TRUE	FALSE
Manual	TRUE	TRUE	TRUE	TRUE

Πίνακας 5.12: Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "Which is the capital of Greece"

When did Socrates die?	URL 1	URL 2	URL 3	URL 4	URL 5
Pairwise Similarity	0,3362579624	0,2475473	0,3176964	0,07171771	0,00440267
Existance	TRUE	TRUE	TRUE	TRUE	TRUE
Meta Existance	TRUE	TRUE	TRUE	FALSE	TRUE
Manual	TRUE	TRUE	TRUE	TRUE	FALSE

Πίνακας 5.13: Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "When did Socrates die?"

Mass of Earth	URL 1	URL 2	URL 3	URL 4	URL 5
Pairwise Similarity	0,455869	0,340556	0,161714	0,552588	0,522073
Existance	TRUE	TRUE	FALSE	TRUE	TRUE
Meta Existance	FALSE	TRUE	FALSE	TRUE	TRUE
Manual	TRUE	TRUE	TRUE	TRUE	TRUE

Πίνακας 5.14: Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "Mass of Earth"

Είναι προφανές ότι και οι 4 μετρικές έχουν παρόμοια αποτελέσματα και δεν υπάρχουν μεγάλες αποκλίσεις στις επιδόσεις τους. Αδιαμφισβήτητα, το pairwise similarity έχει προοπτικές για περισσότερη πληροφορία καθώς δεν είναι Boolean μεταβλητή. Ωστόσο, υπάρχει η υποψία πως οι τιμές που επιστρέφει ο αλγόριθμος δεν είναι αρκετά αντιπροσωπευτικές. Σε ορισμένες περιπτώσεις οι τιμές ίσως επηρεάζονται από λέξεις του ερωτήματος που δεν είναι ο στόχος, όπως συνδετικές λέξεις κτλ, με αποτέλεσμα να αξιολογούνται υψηλά ιστοσελίδες που θα έπρεπε να απορριφθούν. Στα 14 ερωτήματα που έγιναν, επιστράφηκαν 85 ιστοσελίδες. Από αυτές, μόνο σε 5 η μετρική pairwise similarity ήταν υψηλή ενώ η μεταβλητή Manual ήταν Ψευδής. Η μεταβλητή Existance ήταν ανακριβής σε 16 περιπτώσεις ενώ η μεταβλητή Meta Existance σε 35. Αυτό μας επιστρέφει ποσοστά επιτυχίας 83,5%, 81,2% και 58,8% αντίστοιχα. Όπως ήταν αναμενόμενο, η μεταβλητή Meta Existance, η μεταβλητή που αξιολογεί τα μεταδεδομένα, έχει και τη μικρότερη ακρίβεια. Ωστόσο, σε ερωτήματα που η πληροφορία είναι άφθονη, αυτό δεν θα είναι πρόβλημα αλλά ίσως είναι και πλεονέκτημα. Ο λόγος είναι ότι η αποτυχία αυτής της μεθόδου απορρίπτει αποτελέσματα που δεν θα έπρεπε, κάτι το οποίο μειώνει τον χρόνο εκτέλεσης του αλγορίθμου.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συμπεράσματα

Η παρούσα εργασία πραγματοποιείται την ανάπτυξη ενός Δυναμικού Ψηφιακού Βοηθού Ερωταπαντήσεων από Ιστοσελίδες. Το σύστημα βασίστηκε σε έναν επαναληπτικό αλγόριθμο ο οποίος χρησιμοποιεί Question and Answer μοντέλα για την εξαγωγή των απαντήσεων δυναμικά από το διαδίκτυο.

Τα μοντέλα που χρησιμοποιήθηκαν, 3 Αγγλικά και 3 Πολυγλωσσικά, συγκρίθηκαν σε διάφορες ερωτήσεις και καταστάσεις. Αξιολογήθηκαν ως προς την ποιότητα των απαντήσεών τους και ως προς την χρονική διάρκεια εκτέλεσης του αλγορίθμου. Οι ερωτήσεις ήταν σε διαφορετικές γλώσσες, με το σύστημα να αντιδρά αντίστοιχα. Αποδείχθηκε πως, όπως ήταν αναμενόμενο, το μεγαλύτερο μοντέλο είναι και το πιο χρονοβόρο, αλλά τα αποτελέσματά του ήταν εξαιρετικά ακριβή. Τα μικρότερα μοντέλα, όντας πιο ευέλικτα, έκαναν το σύστημα βιώσιμο χρονικά, θυσιάζοντας ποιότητα και ακρίβεια στις απαντήσεις. Ωστόσο, θεωρήθηκε πως η πώση αυτή είναι αποδεκτό αντάλλαγμα, καθώς πλέον οι χρόνοι εκτέλεσης οδηγούν το σύστημα σε αποτέλεσμα σε ένα εύλογο χρονικό διάστημα.

Παράλληλα, αναπτύχθηκαν μέθοδοι επιλογής ιστοσελίδων έπειτα από αξιολόγησή τους ως προς την εγγύτητα με το ερώτημα του χρήστη. Έγινε χρήση διαφορετικών τεχνικών από την βιβλιογραφία και μη. Οι μέθοδοι, είτε με χρήση γραμμικής άλγεβρας και μετασχηματισμών είτε με την χρήση μεταδεδομένων, έδειξαν ενδιαφέροντα αποτελέσματα. Συγκεκριμένα, για μια ακόμα φορά η ποιότητα και η ταχύτητα βρέθηκαν σε μια κατάσταση όπου τίθεται το δίλημμα ενός συμβιβασμού.

Το αποτέλεσμα είναι ένα σύστημα το οποίο, παρ' ότι δεν είναι σε θέση να ανταγωνιστεί τους σύγχρονους ψηφιακούς βοηθούς ή τα μεγάλα γλωσσικά μοντέλα ως προς την λεπτομέρεια και την αναλυτικότητα στις απαντήσεις του, παρέχει ακριβείς και γρήγορες απαντήσεις στον χρήστη. Η δυνατότητα εξαγωγής της βάσης δεδομένων που προκύπτει από τα ερωτήματα του χρήστη, αποτελεί ένα εργαλείο για την αξιοποίηση του συστήματος σε ένα πιο ιδιωτικό περιβάλλον, ενώ ταυτόχρονα οι απαντήσεις προέρχονται από ένα πιο προσωποποιημένο σύνολο πληροφορίας το οποίο δυναμικά ο χρήστης θα έχει συγκεντρώσει. Μάλιστα, η φορητότητα του συστήματος επιτρέπει και Plug & Play χρήση, καθώς χωρίς καθόλου ρύθμιση από τον χρήστη μπορεί να δεχθεί οποιαδήποτε βάση δεδομένων. Τέλος, ο δυναμικός χαρακτήρας του επιτρέπει στον χρήστη να έχει πρόσβαση σε όλη την πληροφορία που είναι διαθέσιμη στο διαδίκτυο χωρίς περιορισμούς. Σε αντίθεση με τα σύγχρονα προεκπαιδευμένα μοντέλα που κατά κόρον χρησιμοποιούνται από τον μέσο χρήστη, η δυναμικότητα αυτού του συστήματος το καθιστά πάντα ενημερωμένο. Ο χρήστης πλέον μπορεί να έχει έναν ψηφιακό βοηθό εντός και εκτός σύνδεσης, που χρησιμοποιεί δεδομένα είτε δικά του είτε από οποιαδήποτε τμήμα του διαδικτύου.

6.2 Μελλοντικές προεκτάσεις

Όπως είναι προφανές, η υλοποίηση του παρόντος συστήματος αφήνει περιθώρια για βελτίωση. Οι μεγαλύτερες ελλείψεις εντοπίζονται στην ποιότητα των απαντήσεων, ενώ το γραφικό περιβάλλον είναι φτωχό. Επιπρόσθετα, η μεταφορά της εφαρμογής σε αυτοτελείς πλατφόρμες είναι κάτι που θα μπορούσε να σχεδιαστεί.

Όσον αφορά την ποιότητα των απαντήσεων, αυτές είναι συνήθως μονολεκτικές και το σύστημα δεν έχει συνομιλητικό χαρακτήρα. Ορμώμενοι από τα σύγχρονα LLMs, αλλά και τους καθημερινούς ψηφιακούς βοηθούς, η συγκεκριμένη προσέγγιση υστερεί σε φιλικότητα και σε αμεσότητα προς τον χρήστη. Από μια άλλη οπτική πάλι, αυτή η απλότητα στην απάντηση παρέχει μια γρήγορη και κατανοητή απόκριση στο κάθε ερώτημα. Ταυτόχρονα όμως, αδυνατούν αυτά τα μοντέλα να απαντήσουν σε πιο σύνθετα ερωτήματα με ακρίβεια, καθώς μια ακριβής απάντηση σε πιο διευρυμένες ερωτήσεις απαιτεί λεπτομέρειες.

Το γραφικό περιβάλλον δεν πληρεί τα σύγχρονα πρότυπα για μοντέρνο σχεδιασμό, αλλά και για φιλικό ως προς τον χρήστη περιβάλλον, αλλά αυτός δεν ήταν ο σκοπός της συγκεκριμένης διπλωματικής εργασίας. Αντιθέτως, η απλοϊκή απάντηση που επιστρέφεται στον χρήστη θα μπορούσε να αντικαταστήσει την υλοποίηση συστήματος αναζήτησης μέσα σε μια ιστοσελίδα. Είναι σύνθετες οι χρήστες να αναζητούν συγκεκριμένα άρθρα ή υλικό μέσα σε κάποια ιστοσελίδα χρησιμοποιώντας ένα ερώτημα που οι ίδιοι έχουν. Σε αυτό το σημείο θα μπορούσε αυτό το σύστημα να προσαρμοστεί καλύπτοντας αυτήν την ανάγκη. Μια πιθανή προσέγγιση είναι η εξαγωγή μιας ολόκληρης ιστοσελίδας ανά τακτά χρονικά διαστήματα και η χρήση αυτών των πηγών για την εξυπηρέτηση του ερωτήματος του χρήστη. Αυτή η *a priori* γνώση μπορεί να αποτελέσει μια γρήγορη πηγή για την απάντηση οποιουδήποτε ερωτήματος. Παραδείγματος χάριν, εάν ο χρήστης βρίσκεται σε μια αθλητική ειδησεογραφική ιστοσελίδα και κάνει ένα ερώτημα σχετικά με την μεταγραφική απόκτηση ενός αθλητή, τότε το σύστημα μπορεί ταχύτατα, χωρίς τον χρόνο εκτέλεσης και την αναμονή που χρειάζεται για την απόκτηση της πληροφορίας, να απαντήσει στον χρήστη παρουσιάζοντας ταυτόχρονα το πλαίσιο αλλά και το άρθρο από το οποίο έγινε η εξαγωγή της απάντησης. Με αυτόν τον τρόπο, θα μπορούσε να αντικατασταθεί η υπηρεσία της αναζήτησης με ένα νέο πρότυπο, όπου το σύστημα και επιστρέφει το άρθρο που αναζητεί ο χρήστης, αλλά και την απάντηση. Μια τέτοια υλοποίηση θα απαιτούσε κάποιου είδους μεταφορά της εφαρμογής σε *add-on* ή *widget* μορφή, όπου θα γίνεται εύκολα εγκατάσταση. Η ανανέωση της βάσης δεδομένων του συστήματος θα μπορούσε να γίνεται είτε καθημερινά, είτε ωριαία, είτε πιο συχνά, πάντα ανάλογα με την ιστοσελίδα, το περιεχόμενό της και τις ανάγκες της.

Η παραπάνω εξέλιξη απαιτεί και μια τεχνική βελτίωση της εφαρμογής. Το σύστημα θα πρέπει, η τουλάχιστον συνίσταται, να μεταφερθεί σε μια *Platform as a service* (PaaS), όπως παραδείγματος χάριν το *Docker*. Αυτή η αναβάθμιση θα έχει άμεσο αντίκτυπο στην ευκολία εγκατάστασης και χρήσης του συστήματος σε διαφορετικά περιβάλλοντα.

Τέλος, υπάρχουν περιθώρια ως προς την χρήση περισσότερων μοντέλων. Νέα μοντέλα σε περισσότερες γλώσσες μπορούν να εναλλάσσονται δυναμικά δίνοντας στον χρήστη την σωστή απάντηση στην γλώσσα που επιθυμεί. Ταυτόχρονα, διαφορετικοί *Readers* μπορεί να είναι πιο αποτελεσματικοί με διαφορετικά μοντέλα και διαφορετικές γλώσσες επιτυγχάνοντας ταχύτερες αποκρίσεις και καλύτερες απαντήσεις.

Παράρτημα Α΄

Ακρωνύμια και συντομογραφίες

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

BOW Bag-Of-Words

CDF Cumulative Distribution Function

CSS Cascading Style Sheets

CV Computer Vision

DL Deep Learning

DMN Dynamic Memory Network

GB Giga Byte

GPT Generative Pre-trained Transformer

HTML HyperText Markup Language

HTTP Hypertext Transfer Protocol

IBM International Business Machines

ipynb IPython Notebook

JSON JavaScript Object Notation,

LLM Large Language Model

LSA Latent Semantic Analysis

ML Machine Learning

MLM Masked Language Modeling

NLP Natural Language Processing

NLT Natural Language Toolkit

NLU Natural Language Understanding

PaaS Platform as a Service

PDF Probability Density Function

QA Question and Answer

REST API Representational State Transfer Application Programming Interface

RNN Recurrent Neural Network

SVD Singular Value Decomposition

TF-IDF Term Frequency - Inverse Document Frequency

UI User Interface

URL Uniform Resource Locator

XLN Extra Large Model

XML Extensible Markup Language

XXL Extra Extra Large

W3C World Wide Web Consortium

TN Τεχνητή Νοημοσύνη

List of Figures

1.1	Αποτελέσματα αναζήτησης στο Scopus Από το 1966 μέχρι το 2021 με λέξεις κλειδιά “chatbot” ή “conversation agent” ή “conversational interface” [4]	4
2.1	Τυπική Αρχιτεκτονική ενός chatbot [4]	7
2.2	Το μοντέλο αρχιτεκτονικής Transformer [24]	9
2.3	Αυτοπροσοχή	10
2.4	(αριστερά) Γραμμικό Εσωτερικό Γινόμενο. (δεξιά) Multi-head Αυτοπροσοχή αποτελούμενη από παράλληλα layers [24]	11
2.5	Κωδικοποιητής και Αποκωδικοποιητής [25]	12
2.6	Αρχιτεκτονική δικτύων δυναμικής μνήμης [26]	13
3.1	Τυπική Μορφή ενός συστήματος στο Haystack	16
3.2	Το REST API του Haystack	16
4.1	Διαδικασία Απάντησης	20
4.2	Απάντηση χωρίς αναζήτηση	20
4.3	Απάντηση με το SERP API	21
4.4	Απάντηση με το πρώτο αποτέλεσμα της αναζήτησης στο διαδίκτυο	22
4.5	Απάντηση με όλα τα αποτελέσματα της αναζήτησης στο διαδίκτυο	22
4.6	“What is bitcoin?” Συχνή Ερώτηση	24
4.7	Δομή ενός pipeline [34]	25
4.8	Η μπάρα που δέχεται το ερώτημα του χρήστη και οι επιλογές	28
4.9	Η μπάρα με το μενού επιλογής γλώσσας	28
4.10	Η μπάρα με το μενού επιλογής συγκεκριμένης ιστοσελίδας ως πηγή	29
4.11	Η μπάρα με το μενού επιλογής μοντέλου	29
4.12	Πρώτη πηγή	29
4.13	Πρώτα αποτελέσματα	30
4.14	Συγκεντρωτικά αποτελέσματα	30
4.15	Import	31
4.16	Export	32
5.1	Σύγκριση απαντήσεων στο ερώτημα “Who is the President of the USA”	36
5.2	Σύγκριση απαντήσεων στο ερώτημα “who is Cristiano Ronaldo?”	37
5.3	Κατανομή των σκορ του κάθε μοντέλου με τη χρήση Boxplot	38
5.4	Κατανομή των σκορ του κάθε μοντέλου με τη χρήση ραβδογράμματος	38
5.5	Συνάρτηση Πιθανότητας και Αθροιστικής Πιθανότητας των σκορ	40
5.6	Μέση Κορυφαία Απάντηση	40
5.7	Ενδιάμεση Κορυφαία Απάντηση	41
5.8	Μέσοι χρόνοι εκτέλεσης κάθε μοντέλου στο 1ο στάδιο	43
5.9	Μέσοι χρόνοι εκτέλεσης κάθε μοντέλου στο 2ο στάδιο	44
5.10	Κατανομή των χρόνων του κάθε μοντέλου στο κάθε στάδιο με τη χρήση ραβδογράμματος	44

List of Tables

5.1	Μετρικές της ερώτησης "Who is the President of the USA"	35
5.2	Μετρικές της ερώτησης "who is Cristiano Ronaldo?"	36
5.3	Μετρικές της ερώτησης "Which is the capital of Greece"	37
5.4	Γενικές Μετρικές	39
5.5	Εκατοστημόρια	39
5.6	Κορυφαία απάντηση ανά μοντέλο	41
5.7	Κορυφαία απάντηση ανά μοντέλο	42
5.8	Πίνακας με μετρικές για τους χρόνους εκτέλεσης κάθε μοντέλου στο 1ο στάδιο	45
5.9	Πίνακας με μετρικές για τους χρόνους εκτέλεσης κάθε μοντέλου στο 2ο στάδιο	45
5.10	Πίνακας με εκατοστημόρια και ακραίες τιμές κάθε μοντέλου στο 1ο στάδιο	45
5.11	Πίνακας με εκατοστημόρια και ακραίες τιμές κάθε μοντέλου στο 2ο στάδιο	45
5.12	Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "Which is the capital of Greece"	46
5.13	Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "When did Socrates die?"	46
5.14	Πίνακας σχετικά με την αξιολόγηση ιστοσελίδων στο ερώτημα "Mass of Earth"	46

Bibliography

- [1] A. M. TURING, "COMPUTING MACHINERY AND INTELLIGENCE," *Mind*, vol. LIX, no. 236, pp. 433–460, 10 1950. [Online]. Available: <https://doi.org/10.1093/mind/LIX.236.433>
- [2] H. Shah, K. Warwick, J. Vallverdú, and D. Wu, "Can machines talk? comparison of eliza with modern dialogue systems," *Computers in Human Behavior*, vol. 58, pp. 278–295, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563216300048>
- [3] G. E. Moore, "Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, april 19, 1965, pp.114 ff." *IEEE Solid-State Circuits Society Newsletter*, vol. 11, no. 3, pp. 33–35, 2006.
- [4] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666827020300062>
- [5] "Number of personal assistance robots worldwide from 2020 to 2030 (in millions)." [Online]. Available: <https://www.statista.com/statistics/1259870/personal-assistance-robots-worldwide/>
- [6] 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.
- [7] OpenAI, "Gpt-4 technical report," 2023.
- [8] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," 2023.
- [9] "Oracle." [Online]. Available: <https://www.oracle.com/chatbots/what-is-a-chatbot/>
- [10] "Roberta." [Online]. Available: https://huggingface.co/docs/transformers/model_doc/roberta
- [11] "all-minilm-l6-v2." [Online]. Available: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- [12] "Albert xxlarge v2." [Online]. Available: <https://huggingface.co/albert-xxlarge-v2>
- [13] "xlm-roberta-base." [Online]. Available: <https://huggingface.co/xlm-roberta-base>
- [14] "deberta-v3-base." [Online]. Available: <https://huggingface.co/microsoft/deberta-v3-base>
- [15] "deepset/gelectra-base-germanquads." [Online]. Available: <https://huggingface.co/deepset/gelectra-base-germanquad>
- [16] "Readers haystack by deepset." [Online]. Available: <https://docs.haystack.deepset.ai/docs/reader>
- [17] P. H. Winston, *Artificial intelligence*. Addison-Wesley, 1984.

- [18] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California Management Review*, vol. 61, no. 4, pp. 5–14, 2019. [Online]. Available: <https://doi.org/10.1177/0008125619864925>
- [19] G. F. Luger and W. A. Strublefield, *Artificial intelligence: structures and strategies for complex problem solving*. Redwood City, CA.: Benjamin/Cummings Pub. Co., 1993.
- [20] N. Thome and C. Wolf, "Histoire des réseaux de neurones et du deep learning en traitement des signaux et des images," Apr. 2023, working paper or preprint. [Online]. Available: <https://hal.science/hal-04058482>
- [21] C. D. Manning, "Human language understanding & reasoning," *Daedalus*, 2022.
- [22] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," 2021.
- [23] "Towards data science." [Online]. Available: <https://towardsdatascience.com/transformer-in-cv-bbdb58bf335e>
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.
- [25] "The illustrated transformer." [Online]. Available: <http://jalammar.github.io/illustrated-transformer/>
- [26] G. Yigit and M. F. Amasyali, "Ask me: A question answering system via dynamic memory networks," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2019, pp. 1–5.
- [27] "hubspot." [Online]. Available: <https://www.hubspot.com/products/crm/chatbot-builder>
- [28] D. Kuhlman, *A Python Book: Beginning Python, Advanced Python, and Python Exercises*. Platypus Global Media, 2011. [Online]. Available: <https://books.google.gr/books?id=1FL-ygAACAAJ>
- [29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <https://www.tensorflow.org/>
- [30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [32] "Haystack, an open-source llm framework to build production-ready applications." [Online]. Available: <https://haystack.deepset.ai/>
- [33] "elasticsearch." [Online]. Available: <https://www.elastic.co/what-is/elasticsearch>
- [34] "Towards data science." [Online]. Available: <https://ashokpalivela.medium.com/question-answering-system-nlp-project-intermediate-46192a240799>
- [35] 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.
- [36] 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.

- [37] 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.
- [38] R. Qu, Y. Fang, W. Bai, and Y. Jiang, “Computing semantic similarity based on novel models of semantic representation using wikipedia,” *Information Processing & Management*, vol. 54, no. 6, pp. 1002–1021, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306457317309226>
- [39] T. Kenter and M. de Rijke, “Short text similarity with word embeddings,” in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1411–1420. [Online]. Available: <https://doi.org/10.1145/2806416.2806475>
- [40] M. C. Lee, J. W. Chang, and T. C. Hsieh, “A grammar-based semantic similarity algorithm for natural language sentences,” *The Scientific World Journal*, vol. 2014, p. 437162, 2014.
- [41] V. Rawte, A. Gupta, and M. J. Zaki, “A comparative analysis of temporal long text similarity: Application to financial documents,” in *Mining Data for Financial Applications*, V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, G. Ponti, and L. Severini, Eds. Cham: Springer International Publishing, 2021, pp. 77–91.
- [42] “Haystack reader documentation.” [Online]. Available: <https://docs.haystack.deepset.ai/docs/reader>
- [43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” 2020.
- [44] P. He, J. Gao, and W. Chen, “Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing,” 2023.
- [45] T. Möller, J. Risch, and M. Pietsch, “Germanquad and germandpr: Improving non-english question answering and passage retrieval,” 2021.