Προγραμματιστική Εργασία

Φάση 3 Υβριδικό σύστημα ανάκτησης:

συνδυασμός κλασικής και σημασιολογικής ανάκτησης

Συστήματα Ανάκτησης Πληροφοριών Ακαδημαϊκό

Έτος 2025-2026

ΦΑΤΣΕΑ ΑΝΘΙΠΠΗ 3190209

1. Εισαγωγή

Στην τρίτη και τελική φάση της εργασίας, αναπτύχθηκε ένα **Υβριδικό Σύστημα Ανάκτησης**, το οποίο ακολουθεί την αρχιτεκτονική **"Retrieve and Re-rank"** (Ανάκτηση και Επανακατάταξη). Στόχος της υλοποίησης ήταν η αντιμετώπιση των περιορισμών που εμφάνισαν οι μεμονωμένες προσεγγίσεις των προηγούμενων φάσεων:

1. Του **Elasticsearch (BM25)**, το οποίο αν και γρήγορο, βασίζεται σε λεξική ταύτιση (lexical matching) και χάνει το σημασιολογικό πλαίσιο.
2. Της **Dense Retrieval (Embeddings)**, η οποία ενώ κατανοεί τη σημασιολογία, όταν εφαρμόζεται σε ολόκληρη τη συλλογή είναι υπολογιστικά ακριβή και επιρρεπής σε θόρυβο (False Positives).

Η υλοποίηση βασίστηκε στη θεωρία των **Neural Re-Ranking Models** , συνδυάζοντας την ταχύτητα του BM25 για το αρχικό φιλτράρισμα (First stage ranker) με την ακρίβεια των Transformers για την τελική κατάταξη (Second stage re-ranker).

2. Μεθοδολογία και Αρχιτεκτονική Pipeline

Η διαδικασία υλοποιήθηκε μέσω της συνάρτησης *run_pipeline()* και ακολουθεί τρία διακριτά στάδια επεξεργασίας για κάθε ερώτηση.

2.1 Στρατηγική Διπλής Προεπεξεργασίας & Διαχείριση Δεδομένων

Ένα από τα κρισιμότερα τεχνικά ζητήματα που επιλύθηκαν ήταν η ανάγκη για διαφορετική μορφή κειμένου σε κάθε στάδιο. Για το λόγο αυτό, υιοθετήθηκε μια **Στρατηγική Διπλής Προεπεξεργασίας**:

Phase 1 Cleaning (Aggressive Normalization): Για την αρχική ανάκτηση με το Elasticsearch, εφαρμόστηκε "επιθετικός" καθαρισμός *(clean_for_phase1)*. Αυτός περιλαμβάνει **Stemming** (αποκοπή καταλήξεων με PorterStemmer) και αφαίρεση *stopwords/boilerplate* φράσεων. Στόχος είναι η μεγιστοποίηση της **Ανάκλησης**, επιτρέποντας την ταύτιση λέξεων με κοινή ρίζα.

Phase 2 Cleaning (Context Preservation): Αντίθετα, τα μοντέλα Transformers βασίζονται στη φυσική γλώσσα για να εξάγουν νόημα. Εάν τροφοδοτούσαμε το μοντέλο με stemmed κείμενο, θα χανόταν κρίσιμη πληροφορία. Γι' αυτό, χρησιμοποιήθηκε η *clean_for_phase2*, η οποία διατηρεί σημεία στίξης, καταλήξεις και stopwords, πραγματοποιώντας μόνο βασικό καθαρισμό χαρακτήρων HTML.

Για να υποστηριχθεί αυτή η στρατηγική, το σύστημα φορτώνει τα **πρωτότυπα (raw) κείμενα** σε μνήμη *(doc_lookup)* χρησιμοποιώντας τα **IDs** των εγγράφων. Έτσι, ανεξάρτητα από το πώς έχει

αποθηκευτεί το κείμενο στο Index του Elasticsearch, το BERT λαμβάνει πάντα ως είσοδο το πλήρες, φυσικό κείμενο για τη διανυσματοποίηση.

2.2 Τα Στάδια της Ανάκτησης

Αρχική Ανάκτηση (Filtering): Το Elasticsearch εκτελεί ένα σύνθετο **Boolean Query**. Χρησιμοποιείται ο τελεστής *MUST* για την εύρεση λέξεων (Recall) και ο τελεστής *SHOULD* με *match_phrase* και υψηλό *boost* (10.0) για την επιβράβευση ακριβών φράσεων. Ανακτώνται τα **Top-200** έγγραφα.

Σημασιολογική Επανακατάταξη (Re-ranking): Το σύστημα υπολογίζει δυναμικά (on-the-fly) τα embeddings μόνο για τα 200 υποψήφια έγγραφα χρησιμοποιώντας το μοντέλο **all-MiniLM-L6-v2**. Στη συνέχεια, μέσω της βιβλιοθήκης **FAISS**, υπολογίζεται η ομοιότητα συνημιτόνου μεταξύ της ερώτησης και των υποψηφίων.

3. Σύγκριση Στρατηγικών Συγχώνευσης (Fusion Strategies)

Κατά την ανάπτυξη του συστήματος, πειραματιστήκα με δύο μεθόδους για τον συνδυασμό των αποτελεσμάτων του Elasticsearch και των Embeddings:

1. **Γραμμικός Συνδυασμός με Στάθμιση (Weighted Linear Combination):**

Αρχικά, δοκιμάστηκε η προσέγγιση:

$$Score = alpha \cdot BM25_{norm} + beta \cdot Vector.$$

Όπου ***alpha=0.5*** και ***beta=0.5***. Για την υλοποίηση αυτής της μεθόδου, κρίθηκε απαραίτητη η κανονικοποίηση των σκορ του BM25 (τα οποία δεν έχουν άνω φράγμα) στο διάστημα [0,1] μέσω *MinMaxScaler*, ώστε να είναι συγκρίσιμα με τα σκορ ομοιότητας συνημιτόνου (Cosine Similarity) των διανυσμάτων.

Ωστόσο, η μέθοδος αυτή αποδείχθηκε ευαίσθητη σε ακραίες τιμές των σκορ του BM25, απαιτώντας λεπτομερή μικρο-ρύθμιση των παραμέτρων alpha καιbeta για κάθε ερώτημα ξεχωριστά, κάτι που δεν είναι πρακτικό σε πραγματικές συνθήκες.

2. **Reciprocal Rank Fusion (RRF) - (Η επιλεγμένη μέθοδος):**

Τελικά, υιοθετήθηκε ο αλγόριθμος RRF. Σε αντίθεση με τον γραμμικό συνδυασμό, το RRF αγνοεί τα απόλυτα σκορ και βασίζεται αποκλειστικά στη **σειρά κατάταξης (Rank)** των εγγράφων στις δύο λίστες.

$$Score_{RRF} = \frac{1}{k + Rank_{BM25}} + \frac{1}{k + Rank_{Vector}}$$

Η προσέγγιση αυτή αποδείχθηκε πιο στιβαρή, καθώς εξαλείφει την ανάγκη για κανονικοποίηση των σκορ και αντιμετωπίζει αποτελεσματικά το πρόβλημα των διαφορετικών κλιμάκων μεταξύ

των δύο συστημάτων ανάκτησης, οδηγώντας σε σταθερά υψηλότερη απόδοση σε σχέση με την απλή στάθμιση.

4. Πειραματικά Αποτελέσματα

Η αξιολόγηση του συστήματος πραγματοποιήθηκε με το πρότυπο εργαλείο *trec_eval*, χρησιμοποιώντας τις μετρικές **MAP (Mean Average Precision)**, **Precision@k** και **Recall@k**. Επιπλέον, για την οπτική επαλήθευση της απόδοσης, παράχθηκαν καμπύλες **Interpolated Precision-Recall**, οι οποίες αναδεικνύουν τη συμπεριφορά των μεθόδων σε όλο το φάσμα της ανάκλησης.

4.1 Συγκεντρωτικοί Πίνακες Αποτελεσμάτων

Στον **Πίνακα 1** παρουσιάζεται η σύγκριση των τριών φάσεων για το σενάριο $k=50$, το οποίο επιλέχθηκε ως το βέλτιστο καθώς προσφέρει την πληρέστερη εικόνα της ανάκλησης.

Πίνακας 1: Σύγκριση Επιδόσεων Φάσεων (k=50)

Metric	Phase 1	Phase 2	Phase 3
MAP	**0.8137**	0.4361	0.7447
P@5	**0.9200**	0.6600	0.9000
P@10	**0.7900**	0.5400	0.7600
Recall@50	**0.9900**	0.7248	0.9466

Στον **Πίνακα 2** αναλύεται η ευαισθησία της Υβριδικής Μεθόδου σε διαφορετικά βάθη ανάκτησης (k).

Πίνακας 2: Επιδόσεις Υβριδικής Μεθόδου ανά βάθος ανάκτησης (k)

METRIC	K=20	K=30	K=50
MAP	0.6485	0.7138	**0.7447**
PRECISION@5	0.9000	0.9000	0.9000
PRECISION@10	0.7600	0.7600	0.7600
PRECISION@20	0.5700	0.5700	0.5700
RECALL@50	0.7513	0.8678	**0.9466**

4.2 Σχολιασμός Αποτελεσμάτων

Βάσει των παραπάνω μετρήσεων, προκύπτουν τα εξής συμπεράσματα για τη συμπεριφορά του συστήματος

Η Φάση 1 πέτυχε εξαιρετικά υψηλές επιδόσεις (**MAP 0.81**, **P@5 0.92**). Αυτό οφείλεται στην επιθετική βελτιστοποίηση που πραγματοποιήθηκε. Αποδεικνύεται ότι για τη συγκεκριμένη συλλογή δεδομένων, η οποία περιέχει εξειδικευμένη ορολογία, η λεξική ταύτιση είναι εξαιρετικά αποτελεσματική.

Η Φάση 2 υστέρησε σημαντικά (**MAP 0.43**). Αυτό είναι αναμενόμενο για ένα μοντέλο "γενικού σκοπού" (all-MiniLM-L6-v2) που δεν έχει εκπαιδευτεί στα συγκεκριμένα δεδομένα. Το μοντέλο εισήγαγε "θόρυβο", ανακτώντας έγγραφα που ήταν σημασιολογικά κοντά αλλά πρακτικά άσχετα.

Παρόλο που η Φάση 3 δεν ξεπέρασε αριθμητικά τη Φάση 1 (MAP 0.74 vs 0.81), επέδειξε αξιοσημείωτη **στιβαρότητα**. Ο αλγόριθμος **RRF** κατάφερε να "απορροφήσει" το μεγάλο σφάλμα της Φάσης 2. Αντί να καταρρεύσει η απόδοση προς το μέσο όρο των δύο μεθόδων (που θα ήταν ≈0.62), διατήρησε την απόδοση πολύ κοντά στα υψηλά επίπεδα της Φάσης 1. Το **Precision@5** παρέμεινε εξαιρετικά υψηλό (**0.90**), διασφαλίζοντας ότι ο χρήστης λαμβάνει ποιοτικά αποτελέσματα στην πρώτη σελίδα.

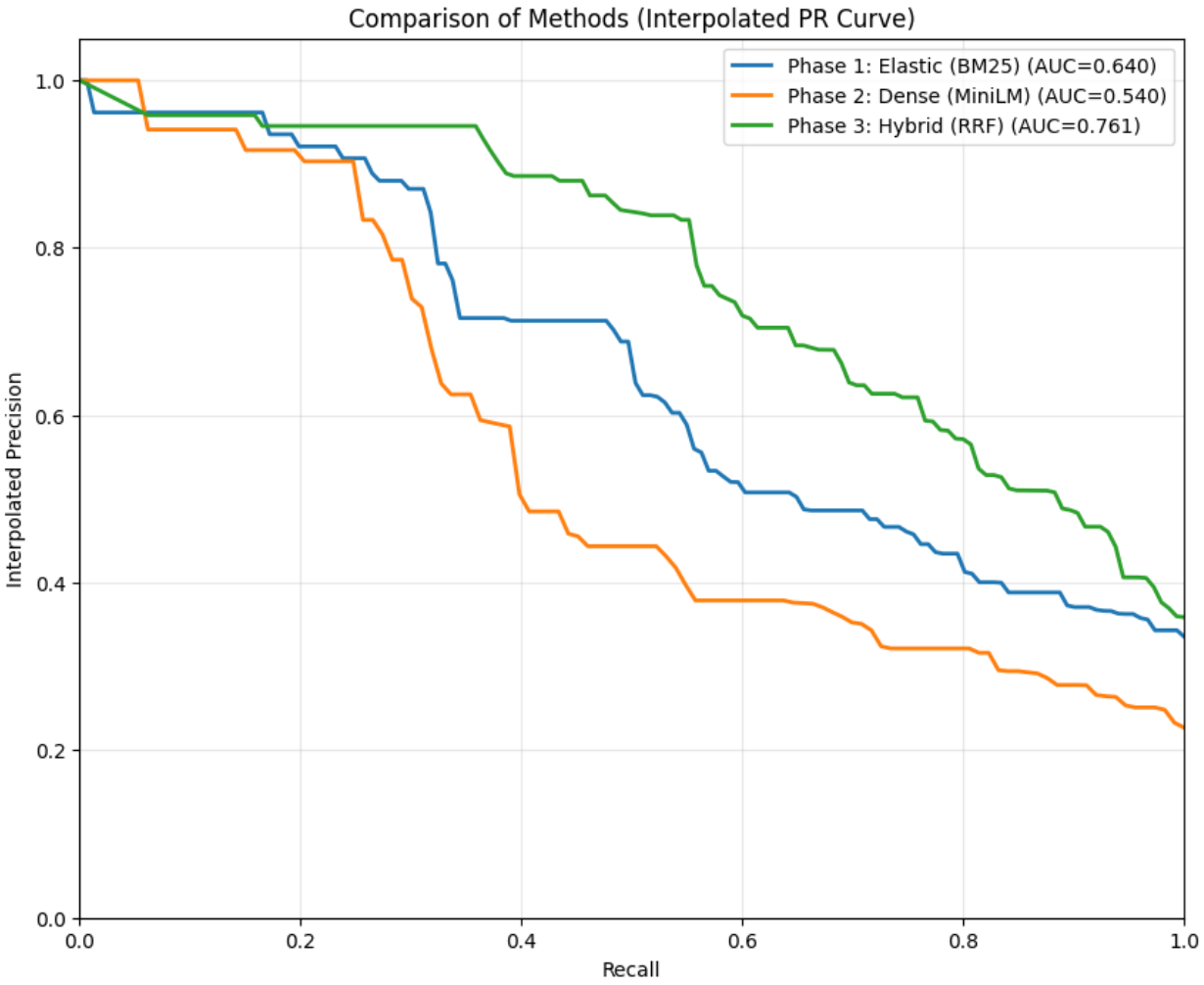
4.3 Ανάλυση Καμπυλών Precision-Recall (Interpolated)

Για την οπτική αξιολόγηση χρησιμοποιήθηκαν καμπύλες **Interpolated Precision-Recall**, ώστε να εξομαλυνθούν οι τοπικές διακυμάνσεις και να αποτυπωθεί η πραγματική τάση των συστημάτων.

Phase 1: Η καμπύλη ξεκινά υψηλά αλλά εμφανίζει αστάθεια και πτώση σε υψηλότερα επίπεδα ανάκλησης, καθώς λεξικά "σωστά" έγγραφα ήταν συχνά σημασιολογικά άσχετα.

Phase 2: Είχε τη χαμηλότερη απόδοση και την πιο βυθισμένη καμπύλη, λόγω του μεγάλου χώρου αναζήτησης και της αδυναμίας διαχείρισης ακριβούς ορολογίας.

Phase 3: Πέτυχε το υψηλότερο **AUC (Area Under Curve = 0.761)**. Η καμπύλη της ξεκινά από το **1.0 (Precision)** και παραμένει σταθερά η πιο ομαλή και υψηλή, επιβεβαιώνοντας ότι ο συνδυασμός των μεθόδων (RRF) προσφέρει την πιο αξιόπιστη συμπεριφορά συνολικά.



5. Συμπεράσματα

Το σύστημα πέτυχε να συνδυάσει την υψηλή ανάκληση του Elasticsearch με τη σημασιολογική πληροφορία του Neural Re-ranking. Παρότι το γενικό μοντέλο της Φάσης 2 εισήγαγε θόρυβο (μειώνοντας ελαφρώς το τελικό MAP σε σχέση με το Baseline), η υβριδική μέθοδος διατήρησε την απόδοση σε υψηλά επίπεδα (0.74), αποφεύγοντας την κατάρρευση που παρατηρήθηκε στην αμιγώς διανυσματική αναζήτηση (0.43). Αυτό επιβεβαιώνει ότι η αρχιτεκτονική 'Retrieve & Re-rank' είναι η πλέον ασφαλής επιλογή για σενάρια όπου η ποιότητα του σημασιολογικού μοντέλου μπορεί να ποικίλλει.