Προγραμματιστική Εργασία

Φάση 2 - Σημασιολογική ανάκτηση χρησιμοποιώντας Transformers και FAISS

Συστήματα Ανάκτησης Πληροφοριών Ακαδημαϊκό
Έτος 2025-2026

ΦΑΤΣΕΑ ΑΝΘΙΠΠΗ 3190209

1. Εισαγωγή

Η παρούσα τεχνική αναφορά εστιάζει στον σχεδιασμό και την υλοποίηση ενός συστήματος **Σημασιολογικής Ανάκτησης** *(Semantic Retrieval)*, στο πλαίσιο της δεύτερης φάσης της εργασίας για το μάθημα "Συστήματα Ανάκτησης Πληροφοριών". Ενώ η προγενέστερη μελέτη μας βασίστηκε σε παραδοσιακές μεθόδους λεξιλογικής ταύτισης **(BM25)**, ανέδειξε σημαντικούς περιορισμούς στη διαχείριση συνωνύμων και πολύπλοκων εννοιών. Στην παρούσα φάση μεταβαίνουμε στη χρήση προηγμένων τεχνολογιών *Νευρωνικών Δικτύων* **(Transformers)**.

Κύριος στόχος της εργασίας είναι η εκμετάλλευση των Πυκνών Διανυσματικών Αναπαραστάσεων (Dense Embeddings) για την "κατανόηση" του νοήματος των επιστημονικών κειμένων, πέρα από την απλή επιφανειακή ταύτιση λέξεων. Χρησιμοποιώντας μοντέλα της οικογένειας **BERT** και βιβλιοθήκες βελτιστοποιημένης αναζήτησης (**FAISS**), αναπτύξαμε ένα σύστημα ικανό να εντοπίζει τη σημασιολογική συγγένεια μεταξύ ερωτημάτων και εγγράφων, ακόμη και όταν δεν υπάρχει κοινό λεξιλόγιο.

Η δομή της αναφοράς αναλύει τη μεθοδολογία και τα ευρήματά μας. Παρουσιάζεται η αρχιτεκτονική του συστήματος, εξηγώντας τη μετάβαση από τον απλό καθαρισμό κειμένου στη δημιουργία διανυσμάτων και την ευρετηρίαση. Περιγράφεται η πειραματική διαδικασία για την επιλογή του βέλτιστου μοντέλου (μεταξύ των **MiniLM**, **MPNet**, **DistilBERT**) και οι τεχνικές βελτιστοποίησης που εφαρμόστηκαν *(Query Pruning)*. Παρατίθενται τα αποτελέσματα της αξιολόγησης με βάση τις μετρικές **MAP**, **Precision** και **Recall**, πλαισιωμένα από οπτικοποιήσεις του διανυσματικού χώρου που επιβεβαιώνουν την ποιότητα της μάθησης του μοντέλου. Τέλος, συνοψίζονται τα συμπεράσματα, όπου η νέα σημασιολογική προσέγγιση συγκρίνεται κριτικά με την προηγούμενη λεξιλογική. Μέσα από τη σύγκριση, αναδεικνύονται τα διακριτά όρια της κάθε μεθόδου. Ενώ η *κλασική αναζήτηση* διατηρεί την πρωτιά στην ακρίβεια, τα *Embeddings* αποδεικνύονται απαραίτητα για την κατανόηση του πλαισίου και την ανάκτηση εγγράφων με διαφορετική ορολογία, οδηγώντας στο συμπέρασμα ότι η βέλτιστη λύση βρίσκεται στον συνδυασμό τους.

2. Μεθοδολογία & Αρχιτεκτονική Συστήματος

Η ποιότητα της σημασιολογικής ανάκτησης εξαρτάται από την ικανότητα του μοντέλου να παράγει ποιοτικά διανύσματα. Αναγνωρίζοντας ότι η ποιότητα της εξόδου καθορίζεται από την ποιότητα της εισόδου, **εστιάσαμε στη προεπεξεργασία,** αντί της άκριτης αφαίρεσης χαρακτήρων, εφαρμόσαμε μια **"Context-Aware" στρατηγική καθαρισμού**, διατηρώντας τεχνικά σύμβολα και δομές που επιτρέπουν στο μοντέλο να διακρίνει λεπτές εννοιολογικές διαφορές, μεγιστοποιώντας έτσι την αντιπροσωπευτικότητα των παραγόμενων διανυσμάτων.

2.1. Ροή Εργασιών

Η λειτουργία του συστήματος αποτελείται από τα εξής διακριτά στάδια:

1. **Φόρτωση & Καθαρισμός:** Εισαγωγή των δεδομένων και σημασιολογικός καθαρισμός.
2. **Διανυσματοποίηση:** Μετατροπή κειμένου σε διανύσματα μέσω Transformer.
3. **Ευρετηρίαση:** Κατασκευή ευρετηρίου αναζήτησης με FAISS.
4. **Αναζήτηση & Αξιολόγηση:** Εκτέλεση ερωτημάτων και υπολογισμός μετρικών *(TREC)*.

2.2. Προεπεξεργασία Δεδομένων (Preprocessing)

Σε αντίθεση με την αφαίρεση όρων *(stemming, αφαίρεση stopword)* της πρώτης φάσης, εδώ ακολουθήθηκε μια "συντηρητική" προσέγγιση. Τα μοντέλα **Transformers** βασίζονται στη σειρά των λέξεων και στα stopwords για να εξάγουν το νόημα. Η αφαίρεσή τους θα κατέστρεφε τη συντακτική δομή που χρειάζεται το μοντέλο.

Εφαρμόσαμε τα εξής βήματα:

- **HTML Decoding:** Μετατροπή HTML entities (π.χ. & → & και &gt → >) για να διατηρηθεί το νόημα σε εκφράσεις όπως "R&D" ή συγκρίσεις τιμών.
- **Normalization (NFKC):** Κανονικοποίηση των χαρακτήρων unicode για ομοιομορφία.
- **Διατήρηση Ειδικών Συμβόλων:** Σε επιστημονικά κείμενα πληροφορικής, σύμβολα όπως `C++`, `R&D`, `.NET`, `O(n)` φέρουν τεράστια σημασία. Αντί να αφαιρέσουμε όλα τα σύμβολα, δημιουργήσαμε μια λίστα επιτρεπόμενων χαρακτήρων (π.χ. `$`, `+`, `&`, `%`) μέσω **Regular Expressions**, ώστε να μην χαθεί η τεχνική πληροφορία.
- **Lowercasing:** Μετατροπή σε πεζά για ομοιομορφία.

2.3. Διανυσματοποίηση (Vectorization)

Ο πυρήνας του συστήματος είναι η μετατροπή των εγγράφων d και των ερωτημάτων q σε διανύσματα v_d, $v_q \in \mathbb{R}^n$. Χρησιμοποιήσαμε τη βιβλιοθήκη **sentence-transformers** η οποία παράγει **embeddings** ειδικά εκπαιδευμένα για τη σύγκριση προτάσεων.

- **Κανονικοποίηση (Normalization):** Όλα τα διανύσματα κανονικοποιήθηκαν σε μοναδιαίο μήκος ($||v|| = 1$). Αυτό επιτρέπει στον υπολογισμό του Εσωτερικού Γινομένου να ταυτίζεται με την Ομοιότητα Συνημιτόνου, η οποία είναι το standard μέτρο σύγκρισης σημασιολογικής εγγύτητας.
- Το κύριο μοντέλο που επιλέχθηκε μετά από πειραματισμό ήταν το **all-MiniLM-L6 v2**.

2.4. Ευρετηρίαση & Αναζήτηση

Για την αποδοτική αναζήτηση στον διανυσματικό χώρο, χρησιμοποιήθηκε η βιβλιοθήκη **FAISS (Facebook AI Similarity Search)**. Επιλέχθηκε ο δείκτης **IndexFlatIP**, ο οποίος εκτελεί **Exact Search** χρησιμοποιώντας εσωτερικό γινόμενο. Δεδομένου του μεγέθους της συλλογής, η χρήση Brute Force είναι εφικτή και εγγυάται ότι θα βρούμε τους πραγματικούς πλησιέστερους γείτονες **χωρίς την απώλεια ακρίβειας** που θα εισήγαγαν προσεγγιστικοί αλγόριθμοι.

3. Πειραματική Διαδικασία

Για την επιλογή του βέλτιστου μοντέλου, διεξήχθησαν πειράματα με τέσσερις διαφορετικές αρχιτεκτονικές **Transformers**. Αξιολογήθηκαν με βάση τη μετρική **MAP** (Mean Average Precision) και την **Ακρίβεια** και την **Ανάκληση** (Recall).

3.1. Διαχείριση Μήκους Εισόδου

Κατά τη φάση ανάπτυξης διερευνήσαμε πώς το μήκος του κειμένου εισόδου επηρεάζει την απόδοση του μοντέλου SBERT.

Αρχικά εξετάστηκε η βιβλιογραφική προσέγγιση του **Chunking**. Ωστόσο, διαπιστώθηκε ότι για το συγκεκριμένο dataset *(Scientific Abstracts)*, η προεπιλεγμένη στρατηγική του μοντέλου *(Truncation στα 512 tokens)* είναι η βέλτιστη, καθώς η ουσιώδης πληροφορία βρίσκεται σχεδόν πάντα στην αρχή της περίληψης.

Υλοποιήθηκε μια πειραματική συνάρτηση **shorten_query**, η οποία εφάρμοζε αυστηρή περικοπή των ερωτήσεων στις 50 πρώτες λέξεις.

```
def shorten_query(text):

    text = clean_text(text)
    words = text.split()
    if len(words) > 50:
        return " ".join(words[:50])
    return text
```

Η σκέψη ήταν ότι οι πολύ μεγάλες ερωτήσεις εισάγουν **"θόρυβο"** και μπερδεύουν το διάνυσμα.

Τα πειράματα έδειξαν ότι η αυστηρή περικοπή αφαιρούσε συχνά **κρίσιμες** λεπτομέρειες από το τέλος της ερώτησης. Στην τελική υλοποίηση, επιλέχθηκε να αφαιρεθεί ο χειροκίνητος περιορισμός των 50 λέξεων και να αφεθεί το μοντέλο να διαχειριστεί ολόκληρο το **context**, καθώς ο μηχανισμός **Attention** αποδείχθηκε ικανός να φιλτράρει το θόρυβο αυτόματα.

Αν και η στρατηγική της περικοπής *(truncation)* λειτούργησε ικανοποιητικά για τα επιστημονικά abstracts, η μελλοντική δοκιμή τεχνικών **Document Chunking** *(κατάτμηση κειμένου)* θα μπορούσε να ξεκλειδώσει πληροφορία που βρίσκεται στο τέλος εκτενέστερων κειμένων, βελτιώνοντας περαιτέρω την ανάκληση του συστήματος.

3.2. Πειραματική Διερεύνηση: Στρατηγική Προεπεξεργασίας

Δοκιμάστηκε η κλασική προσέγγιση της Φάσης 1 *(αφαίρεση σημείων στίξης και ειδικών χαρακτήρων)*. Ωστόσο, παρατηρήθηκε ότι αυτή η μέθοδος **αλλοίωνε σημαντικά τη σημασιολογία** σε τεχνικούς όρους.

Για παράδειγμα, όροι όπως *C++*, *R&D*, *>95%* έχαναν την πληροφοριακή τους αξία όταν αφαιρούνταν τα σύμβολα. Μέσω **Regular Expressions** ([^a-zA-Z0-9\s\.,\?!:;@%#\+\-&<>]), επιτρέπουμε ρητά χαρακτήρες όπως οι οποίοι είναι απαραίτητοι για την κατανόηση επιστημονικών μεγεθών και ορολογίας πληροφορικής. Διατηρήθηκαν τα σημεία στίξης, καθώς τα μοντέλα **Transformers** (SBERT) βασίζονται στη συντακτική δομή της πρότασης για να υπολογίσουν σωστά τα **Embeddings**.

3.3. Σύγκριση Μοντέλων Transformers

Τα μοντέλα που εξετάστηκαν:

1. **all-MiniLM-L6-v2:** Βελτιστοποιημένο για ταχύτητα και απόδοση.
2. **all-MiniLM-L12-v2:** Η μεγαλύτερη έκδοση του **all-MiniLM,** με περισσότερα layers.
3. **all-mpnet-base-v2:** Ένα μεγαλύτερο και θεωρητικά ισχυρότερο μοντέλο.

Παρακάτω παρουσιάζονται τα συγκεντρωτικά αποτελέσματα για *k=50*, όπως προέκυψαν από το εργαλείο trec_eval:

Μοντέλο	MAP @ 50	P @ 5	P @ 10	Recall @ 50
all-MiniLM-L6-v2	**0.4361**	**0.6600**	**0.5400**	**0.7248**
all-mpnet-base-v2	0.4246	0.6000	0.5100	0.7072
all-mpnet-base-v2 **shorten_query**	0.3865	0.5400	0.4400	0.6800
all-MiniLM-L12-v2	0.3385	0.5200	0.4100	0.6539

Παρατηρήσεις:

- Το μοντέλο **all-MiniLM-L6-v2** πέτυχε την υψηλότερη ακρίβεια (P@5 = 0.66) και το καλύτερο MAP (0.4360), ξεπερνώντας ακόμα και το βαρύτερο all-mpnet-base-v2.
- Τα άλλα μονέλα είχαν σημαντικά χαμηλότερη απόδοση.

Η ανάλυση των αποτελεσμάτων οδηγεί σε σημαντικά συμπεράσματα:

1. Το μοντέλο **all-MiniLM-L6-v2** πέτυχε την υψηλότερη ακρίβεια στα πρώτα 5 αποτελέσματα (**66%**), ξεπερνώντας ακόμα και το πολύ βαρύτερο mpnet-base. Αυτό υποδεικνύει ότι για το συγκεκριμένο domain, η συμπαγής αναπαράσταση του MiniLM είναι πιο αποδοτική και λιγότερο επιρρεπής σε θόρυβο.
2. **Η αποτυχία των παλαιότερων μοντέλων:** Το L12 εμφάνισε χαμηλότερη επιδόση. Αυτό επιβεβαιώνει την ταχεία εξέλιξη στον τομέα των Embeddings.

3.4. Επιλογή Μετρικής Ομοιότητας (IP vs L2)

Κατά τη φάση σχεδιασμού του ευρετηρίου **FAISS**, τέθηκε το ερώτημα εάν η επιλογή της μετρικής απόστασης επηρεάζει την ποιότητα της ανάκτησης.

- **IndexFlatIP**: Μετρική Εσωτερικού Γινομένου (ισοδύναμη με Cosine Similarity σε κανονικοποιημένα διανύσματα).
- **IndexFlatL2**: Μετρική Ευκλείδειας Απόστασης.

Τα αποτελέσματα για το ενδεικτικό ερώτημα **Q01** έδειξαν το εξής ενδιαφέρον φαινόμενο:

RANK	DOC ID	SCORE (COSINE/IP)	SCORE (L2 DERIVED)
1	193378	0.9816	0.9644
2	210137	0.6673	0.6005
3	193715	0.6432	0.5836
4	193373	0.6295	0.5744
5	193722	0.6264	0.5723

Παρόλο που οι απόλυτες τιμές των σκορ διαφέρουν, η σειρά κατάταξης των εγγράφων παραμένει **απολύτως ταυτόσημη**.

Το πείραμα επιβεβαίωσε τη μαθηματική σχέση $|u - v|^2 = 2(1 - \cos(u, v))$ για μοναδιαία διανύσματα.

Δεδομένου ότι η κατάταξη δεν αλλάζει, επιλέχθηκε τελικά ο δείκτης **IndexFlatIP**, καθώς η μέτρηση της **ομοιότητας συνημιτόνου (Cosine Similarity)** μέσω του εσωτερικού γινομένου αποδεικνύεται πιο αποτελεσματική από την Ευκλείδεια απόσταση για την ανάλυση κειμένων.

4. Αποτελέσματα

Τα αποτελέσματα προέκυψαν από τη χρήση του μοντέλου **all-MiniLM-L6-v2**, το οποίο επιλέχθηκε ως η βέλτιστη λύση κατά την πειραματική διαδικασία.

4.1 Ποσοτική Ανάλυση

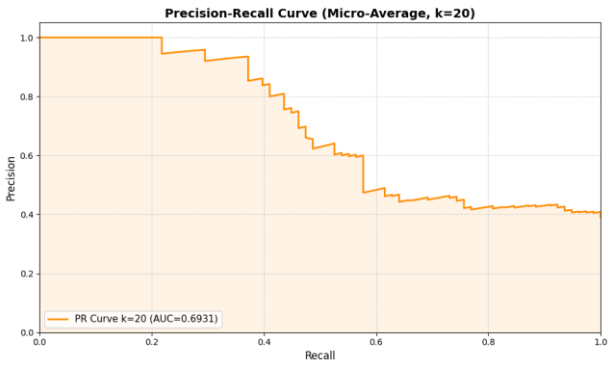
Όπως παρατηρούμε, οι μετρικές ακρίβειας στα πρώτα αποτελέσματα (*P@5*, *P@10*, *P@20*) παραμένουν σταθερές, καθώς οι πρώτες 20 θέσεις της κατάταξης δεν επηρεάζονται από την αύξηση του *k*. Ωστόσο, η αύξηση του *k* σε 50 βελτιώνει σημαντικά το **MAP** από **0.36** σε **0.43** και την Ανάκληση *(R@50)* από **0.49** σε **0.72**, καθώς το σύστημα έχει την ευκαιρία να συμπεριλάβει περισσότερα σχετικά έγγραφα που βρίσκονταν χαμηλότερα στη λίστα.

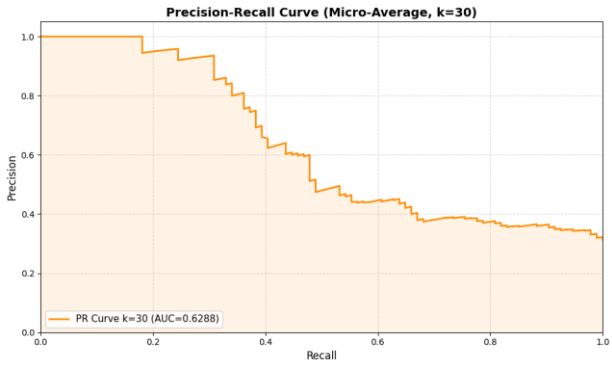
K	MAP	P @ 5	P @ 10	P @ 15	P @ 20	R @ 5	R @ 10	R @ 20	R @ 50
20	0.3612	0.6600	0.5400	0.4600	0.3850	0.2192	0.3465	0.4962	0.4962
30	0.3989	0.6600	0.5400	0.4600	0.3850	0.2192	0.3465	0.4962	0.6022
50	**0.4361**	0.6600	0.5400	0.4600	0.3850	0.2192	0.3465	0.4962	**0.7248**

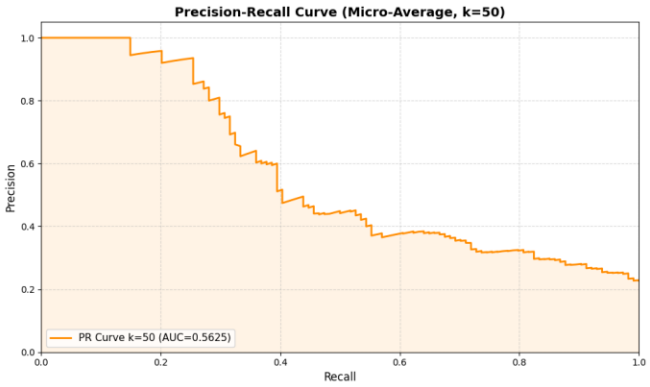
4.2. Καμπύλες Precision-Recall

Η γραφική παράσταση της καμπύλης Precision-Recall για το βέλτιστο μοντέλο **(all-MiniLM-L6-v2)** δείχνει τη συμπεριφορά του συστήματος καθώς αυξάνεται το πλήθος των ανακτώμενων εγγράφων.

Οι καμπύλες ξεκινούν από ψηλά, μετά υπάρχει πτώση της ακρίβειας καθώς αυξάνεται η ανάκληση.

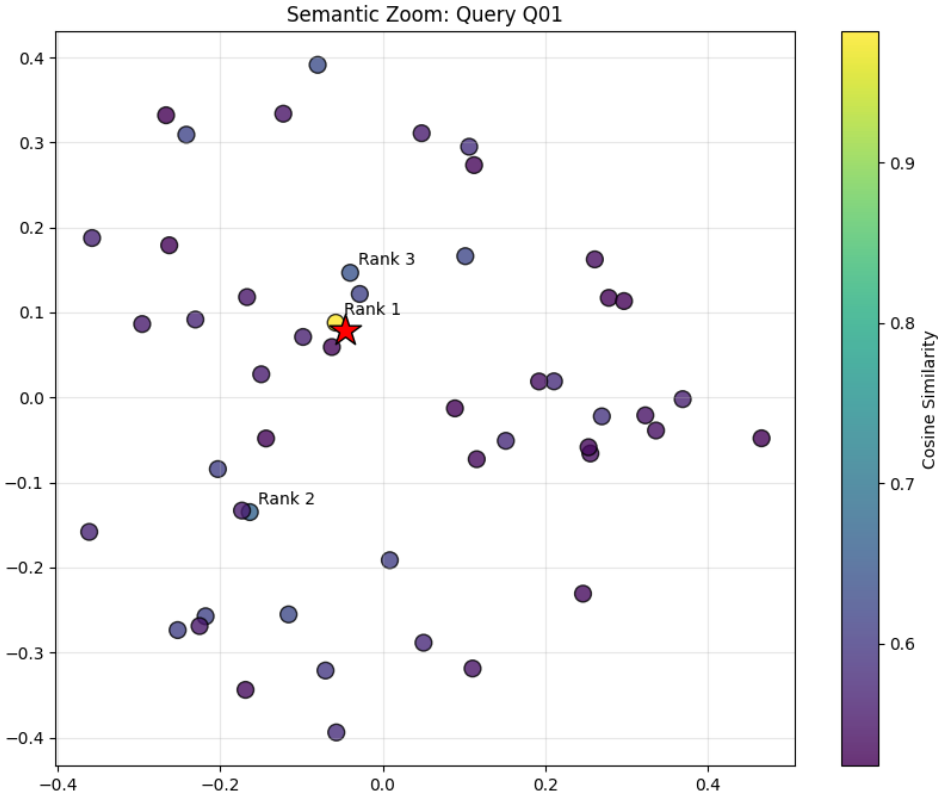




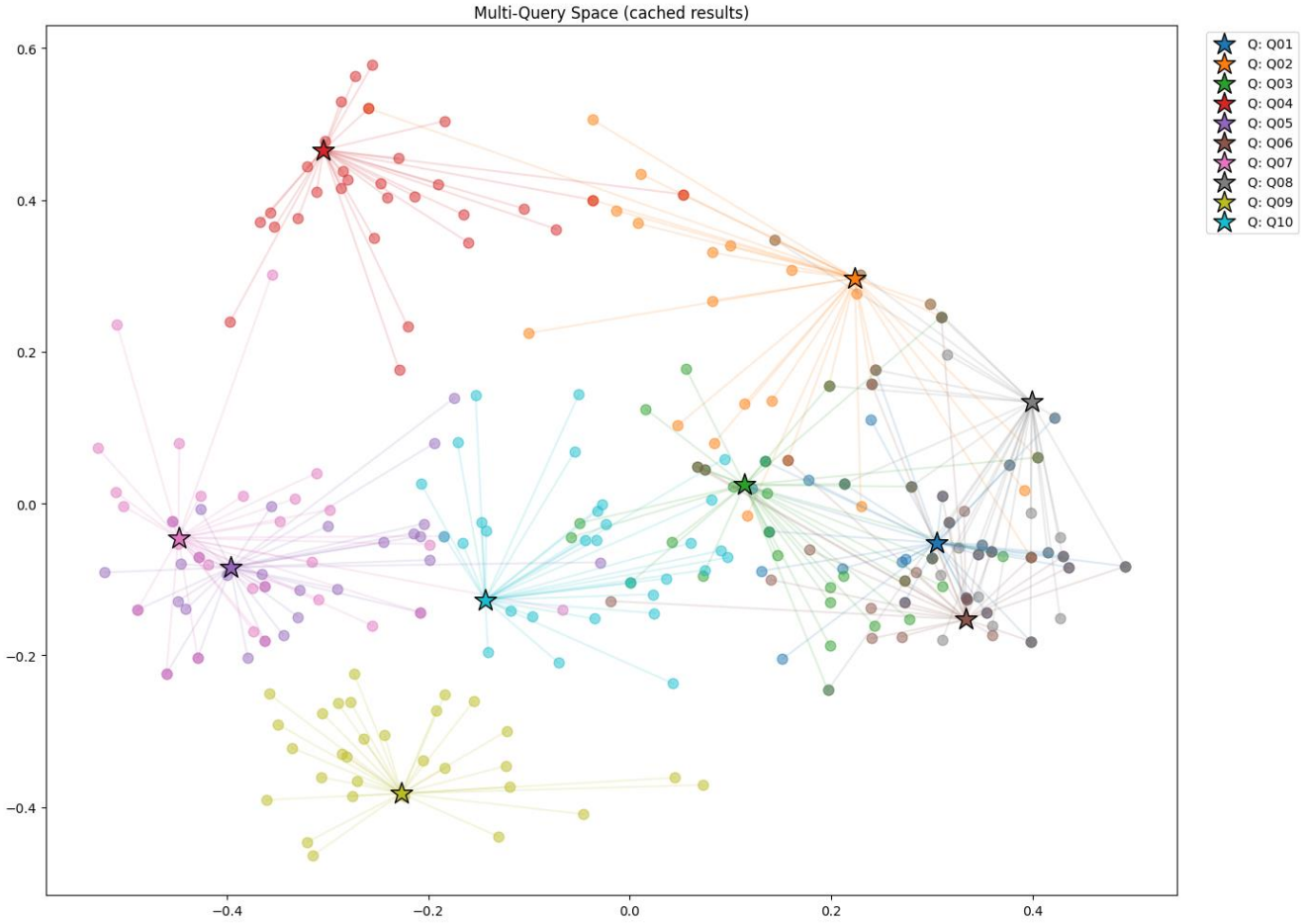


4.3. Οπτικοποίηση Διανυσματικού Χώρου (PCA)

Στο παρακάτω διάγραμμα απεικονίζεται ένα ερώτημα (κόκκινο αστέρι) και τα ανακτώμενα έγγραφα. Παρατηρούμε ότι τα έγγραφα που επισημάνθηκαν ως σχετικά είναι γεωμετρικά πιο κοντά στο ερώτημα, επιβεβαιώνοντας ότι η **Cosine Similarity** αντικατοπτρίζει την πραγματική συνάφεια.



Προβάλλοντας πολλαπλά ερωτήματα ταυτόχρονα, βλέπουμε ότι σχηματίζονται διακριτές συστάδες *(clusters)*. Το σύστημα δεν μπερδεύει τα ερωτήματα μεταξύ τους και κάθε ερώτημα καταλαμβάνει τη δική του "γειτονιά" στον σημασιολογικό χώρο, κάτι που αποδεικνύει την ικανότητα του μοντέλου να διαχωρίζει σωστά τις διαφορετικές θεματικές περιοχές.



4.4. Φάση 1 (BM25) vs Φάση 2 (Embeddings)

Σύγκριση μεταξύ του συστήματος της Φάσης 1 (Elasticsearch/BM25) και του συστήματος της Φάσης 2 (SBERT/FAISS). Η σύγκριση βασίζεται στις καμπύλες Precision-Recall, στις μετρικές για k=50 και στα αποτελέσματα που έδωσαν στην σύγκριση των κείμενων που έδωσαν ως απάντηση στις ερωτήσεις.

4.4.1. Σύγκριση Καμπυλών (PR Curves)

- **Φάση 1 (Lexical Search)**: Η καμπύλη PR παρουσιάζει μια σχεδόν "τετραγωνισμένη" μορφή, διατηρώντας την Ακρίβεια κοντά στο 1.0 για μεγάλο εύρος της Ανάκλησης.
- **Φάση 2 (Semantic Search)**: Η καμπύλη του MiniLM είναι πιο ομαλή και φθίνουσα. Ξεκινάει με υψηλή ακρίβεια, αλλά πέφτει γρηγορότερα.

4.4.2. Ποσοτική Σύγκριση

Μετρική	Φάση 1	Φάση 2	Διαφορά
MAP	0.8137	0.4361	-46%
Precision@5	0.9200	0.6600	-28%
Recall@50	0.9900	0.7248	-27%

4.4.3 Ερμηνεία Διαφορών

Η φαινομενική υπεροχή της Φάσης 1 οφείλεται σε συγκεκριμένους παράγοντες σχεδιασμού:

Στη Φάση 1, το Recall εκτοξεύτηκε στο 99% κυρίως λόγω της χρήσης εκτεταμένων λιστών συνωνύμων (π.χ. *"Cancer" -> "Oncology"*, *"UK" -> "Great Britain"*) και πολύπλοκων **Boolean Queries**. Στη Φάση 2, το μοντέλο κλήθηκε να βρει αυτές τις σχέσεις αυτόματα, χωρίς καμία εξωτερική βοήθεια ή λεξικό.

Τα ερωτήματα της συλλογής περιείχαν πολλούς ειδικούς όρους και κωδικούς έργων. Ο αλγόριθμος **BM25** είναι σχεδιασμένος να τους βρίσκει τέλεια *(Exact Match)*, ενώ τα **Embeddings** μερικές φορές "θολώνουν" τους ακριβείς όρους υπέρ του γενικού νοήματος.

Η Φάση 1 διαχειρίστηκε άγνωστες λέξεις μέσω **Fuzzy Search, Stemming και Συνωνύμων**. Στη Φάση 2, αν μια λέξη δεν υπάρχει στο λεξιλόγιο του BERT, το διάνυσμά της μπορεί να μην είναι ακριβές.

Συμπερασματικά, η Φάση 1 αποδείχθηκε καλύτερη όταν γνωρίζουμε ακριβώς τι ψάχνουμε *(Keyword Search),* ενώ η Φάση 2 *(Embeddings)* έδειξε τη δύναμή της στο να κατανοεί το εννοιολογικό πλαίσιο, λειτουργώντας καλύτερα ως εργαλείο ανακάλυψης παρά ως εργαλείο ακριβούς ανάκτησης.

4.4.4 Case Study: Η Περίπτωση της Ερώτησης Q10

Μια χαρακτηριστική περίπτωση που αναδεικνύει την ποιοτική διαφορά των δύο μεθόδων είναι η Ερώτηση 10.

Στην πρώτη φάση, η συγκεκριμένη ερώτηση σημείωσε **εξαιρετικά χαμηλά** ποσοστά ανάκτησης. Η αιτία εντοπίστηκε στο φαινόμενο της λεξιλογικής αναντιστοιχίας. Οι όροι που χρησιμοποιούσε ο χρήστης στην ερώτηση δεν υπήρχαν αυτούσιοι στα σχετικά έγγραφα, και ούτε η επέκταση συνωνύμων κατάφερε να καλύψει το κενό.

```
QUERY ID: Q10 (Στόχος: 10) | Query Text: netcommons network infrastructure as commons communication and information distribut
ion are key comp...
-----------------------------------------------------------------------------------------------
   1. [6880.1840] MATCH! | DOC: 199879 | network infrastructure as commons...
   2. [1766.8068] MATCH! | DOC: 199849 | a diy networking toolkit for location based collective aware...
   3. [1719.1540] MATCH! | DOC: 194285 | universal mobile centric and opportunistic communications ar...
   4. [1675.8527] MATCH! | DOC: 204439 | resolving the tussle in the internet...
   5. [1669.2871] MATCH! | DOC: 194229 | architecture for an internet for everybody...
   6. [1611.6950] MATCH! | DOC: 200424 | understanding collective awareness platforms with the maker ...
   7. [1539.1958] No...  | DOC: 195865 | innovative coherent detection optical access networks...
   8. [1520.0731] MATCH! | DOC: 198820 | digital social innovation for europe...
   9. [1507.2334] No...  | DOC: 194310 | trustful hyper linked entities in dynamic networks...
   10. [1483.9127] No...  | DOC: 205470 | bioeconomy awareness and discourse project...
```

Αντίθετα, κατά την πειραματική διαδικασία της δεύτερης φάσης, το σύστημα επέδειξε εντυπωσιακή βελτίωση μόνο για τη συγκεκριμένη ερώτηση. Συγκεκριμένα, με τη χρήση του ισχυρού μοντέλου **all-mpnet-base-v2** σε συνδυασμό με την πειραματική τεχνική **Query Pruning** *(shorten_query)*, ανακτήθηκαν με επιτυχία τα σχετικά έγγραφα.

```
Query: 5gex 5g exchange the goal of the 5g exchange 5gex project is to enable cross-domain orchestration of
-----------------------------------------------------------------------------------------------
Rank  | Doc ID          | Score    | Status   | Snippet
-----------------------------------------------------------------------------------------------
1     | 197346          | 0.8789   | REL      | 5g exchange: the goal of the 5g exchange 5gex project is to ...
2     | 198311          | 0.6897   | REL      | application-aware user-centric programmable architectures fo...
3     | 211063          | 0.6829   | REL      | 5g development and validation platform for global industry-s...
4     | 197344          | 0.6828   | NO       | 5g-crosshaul: the 5g integrated fronthaul backhaul: mobile d...
5     | 211083          | 0.6739   | REL      | a holistic, innovative framework for the design, development...
6     | 205594          | 0.6725   | REL      | a network slice for every service: 5g!pagoda represents the ...
7     | 211067          | 0.6651   | REL      | 5g-transformer: 5g mobile transport platform for verticals: ...
8     | 197343          | 0.6589   | REL      | small cells coordination for multi-tenancy and edge services...
9     | 211072          | 0.6567   | REL      | embedded network services for 5g experiences: 5g essence add...
10    | 211091          | 0.6543   | REL      | 5g programmable infrastructure converging disaggregated netw...
```

Το μοντέλο **MPNet**, μπόρεσε να δημιουργήσει διανυσματικές συσχετίσεις μεταξύ των εννοιών της ερώτησης και των εγγράφων, γεφυρώνοντας το χάσμα που άφησε η απλή σύγκριση λέξεων. Το "κόψιμο" της ερώτησης βοήθησε περαιτέρω, αφαιρώντας πιθανό θόρυβο και εστιάζοντας το διάνυσμα στον πυρήνα του νοήματος.

5. Συμπεράσματα

Η εκπόνηση της δεύτερης φάσης της εργασίας επέτρεψε τη βαθύτερη κατανόηση των δυνατοτήτων των **Neural Embeddings** στην ανάκτηση πληροφορίας. Η μετάβαση από τη λεκτική στη σημασιολογική ανάκτηση προσέφερε διακριτά πλεονεκτήματα, αλλά ανέδειξε και σημαντικές προκλήσεις.

Βασικά Ευρήματα:

- Το σύστημα πέτυχε να γεφυρώσει το "λεξιλογικό χάσμα". Η ικανότητα των **Transformers** να αντιλαμβάνονται συνώνυμα και τεχνικές έννοιες επέτρεψε την ανάκτηση σχετικών εγγράφων ακόμη και σε περιπτώσεις μηδενικής λεκτικής ταύτισης (όπως παρατηρήθηκε στο ερώτημα Q10).
- Με **P@5 = 0.66**, επιβεβαιώθηκε ότι το μοντέλο κατατάσσει τα πλέον σχετικά έγγραφα στις πρώτες θέσεις, εξασφαλίζοντας μια ποιοτική εμπειρία αναζήτησης για τον χρήστη.
- Η επιλογή του μοντέλου **MiniLM-L6** σε συνδυασμό με τη βιβλιοθήκη **FAISS** απέδειξε ότι η σημασιολογική αναζήτηση μπορεί να είναι εξαιρετικά ταχεία και κλιμακώσιμη, καθιστώντας την κατάλληλη για παραγωγικά συστήματα πραγματικού χρόνου.

Παρά την ευφυΐα των διανυσματικών μοντέλων, η σύγκριση με την πρώτη φάση (BM25) έδειξε ότι η κλασική αναζήτηση παραμένει ισχυρότερη στην ακριβή ταύτιση τεχνικών όρων και κωδικών. Η πειραματική μας ανάλυση οδηγεί στο συμπέρασμα ότι **η βέλτιστη αρχιτεκτονική δεν είναι η επιλογή μίας εκ των δύο μεθόδων, αλλά ο συνδυασμός τους.**

Ένα υβριδικό σύστημα που συνδυάζει τη γραμμική ακρίβεια του BM25 με τη σημασιολογική εμβάθυνση των Transformers θα μπορούσε να προσφέρει το μέγιστο δυνατό MAP, καλύπτοντας τόσο τις περιπτώσεις αναζήτησης με λέξεις-κλειδιά όσο και τις περιπτώσεις εννοιολογικής διερεύνησης.