



**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ**

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Εργασία Εργαστηρίου

Τσελάνι Μαρίνο

ΑΜ:20390241

Γουρδομιχάλης Δημήτριος

ΑΜ:20390043

ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2023-2024

Εισαγωγή

Στα πλαίσια του μαθήματος Ανάκτηση Πληροφορίας, με βάση την εκφώνηση της εργαστηριακής εργασίας, αναπτύχθηκε μια μηχανή αναζήτησης ακαδημαϊκών εργασιών η οποία έχει συγγραφεί σε γλώσσα προγραμματισμού Python.

Το πρόγραμμα αποτελείται από μια γραφική διεπαφή χρήστη(GUI) στην οποία οι χρήστες έχουν την δυνατότητα να αναζητούν και να ανακτούν τα έγγραφα με πολλαπλούς αλγόριθμους ανάκτησης και φίλτρα. Συγκεκριμένα αναπτύχθηκαν οι αλγόριθμοι ανάκτησης Boolean retrieval, Vector Space Model και Okapi BM25(Probabilistic retrieval model). Ως φίλτρα χρησιμοποιήθηκαν η ημερομηνία δημοσίευσης και ο συγγραφέας.

Η συλλογή των δεδομένων γίνεται από έναν σταχυολογητή (web crawler) από το αποθετήριο ακαδημαϊκών εργασιών arXiv. Στη συνέχεια, τα κειμενικά περιεχόμενα(abstract) που προκύπτουν από το προηγούμενο βήμα προεξεργάζονται. Έπειτα, δημιουργούμε μία ανεστραμμένη δομή δεδομένων ευρετηρίου (inverted index) για την αποτελεσματική αντιστοίχιση όρων στα έγγραφα στα οποία εμφανίζονται. Παράλληλα, υλοποιούμε ένα λεξικό(dictionary) για την αποθήκευση της αντιστοίχισης μεταξύ λέξεων και εγγράφων. Επιπροσθέτως, στο επόμενο βήμα, υλοποιούμε μία απλή διεπαφή ιστού και τους 3 αλγόριθμους ανάκτησης που αναφέραμε παραπάνω. Επιπλέον, δημιουργούμε μία συνάρτηση, ώστε να φιλτράρονται τα αποτελέσματα της ανάκτησης, με βάση ημερομηνία δημοσίευσης του άρθρου ή/και συγγραφέα του άρθρου που έχει ανακτηθεί. Αξίζει να τονιστεί το γεγονός ότι οι χρήστες μπορούν να αναζητούν έγγραφα χρησιμοποιώντας μία ή περισσότερες λέξεις(boolean πράξεις AND/OR/NOT). Τέλος, υλοποιούμε τον αλγόριθμο κατάταξης TF-IDF και τον cosine similarity, ώστε να γίνει κατάταξη των αποτελεσμάτων της ανάκτησης με πολλαπλούς τρόπους.

Στη συνέχεια του συγκεκριμένου αρχείου ακολουθεί η αξιολόγηση του συστήματος(Βήμα 5) και η τεκμηρίωση(Βήμα 6).

Αξιολόγηση συστήματος (Βήμα 5^ο)

Για την αξιολόγηση του συστήματος υπάρχουν διάφορες μεθοδολογίες αλλά στην προκειμένη περίπτωση, τα σύνολα δεδομένων που έχουμε αποκομίσει από την ιστοσελίδα έχουν ταξινομηθεί από τον κώδικα μας (συνάρτηση **rank documents**), επομένως θα κινηθούμε με την μεθοδολογία της Ταξινομημένης Αξιολόγησης και τις μετρικές της. Μερικά από αυτά τα μέτρα είναι η καμπύλη Ακρίβειας-ανάκλησης, MAP, R-precision, η DCG 11-point precision και P@k.

Ακολουθεί μια σύντομη περιγραφή του κάθε μέτρου:

- **Precision-Recall Curve (Καμπύλη Precision-Recall):** Είναι ένα γράφημα που δείχνει το εύρος των τιμών της Precision και της Recall για διάφορα κατώφλια αποφάσεων ταξινόμησης.
- **MAP:** Μέση ακρίβεια (AP) για ένα ερώτημα q με σχετικά έγγραφα $\{d_1, \dots, d_m\}$ είναι ο μέσος όρος των βαθμολογιών ακρίβειας που μετρούνται στις τάξεις των σχετικών εγγράφων, η MAP είναι η AP που υπολογίζεται κατά μέσο όρο στο σύνολο των ερωτημάτων Q ,

$$\text{MAP} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk})$$

- **R-Precision :** Είναι η ακρίβεια στο σημείο όπου ο αριθμός των ανακαλούμενων εγγράφων ισούται με το πλήθος των πραγματικών θετικών εγγράφων.
- **nDCG:** Μετρά την ποιότητα της σειράς των αποτελεσμάτων αναζήτησης, λαμβάνοντας υπόψη τη σημασία των θετικών εγγράφων.
- **11-point precision :** Η ακρίβεια 11 σημείων περιγράφει την απόδοση ενός συστήματος IR μέσω ακρίβειας που μετριέται σε 11 διαφορετικά επίπεδα ανάκλησης.
- **P@k (Precision at k - Ακρίβεια στο k):** Υπολογίζει την ακρίβεια των πρώτων k αποτελεσμάτων στη σειρά.

Σε διαφορετική περίπτωση, η αξιολόγηση του συστήματος θα μπορούσε να υλοποιηθεί με τις μετρικές Ακρίβεια / Ανάκληση / F(F1 score) με τις οποίες έχουμε ασχοληθεί σε θεωρητικό επίπεδο στο μάθημα της Θεωρίας. Πιο αναλυτικά:

- **Ακρίβεια:** είναι το ποσοστό των ανακτηθέντων εγγράφων που είναι συναφή

$$precision = true_positives / (true_positives + false_positives)$$

- **Ανάκληση:** είναι το ποσοστό των συναφών εγγράφων που έχουν ανακτηθεί

$$recall = true_positives / (true_positives + false_negatives)$$

- **F1 Score:** επιτρέπει την εναλλαγή ακρίβειας – ανάκλησης, δηλαδή είναι ο αρμονικός μέσος των δύο, καθώς η ακρίβεια με την ανάκληση, αλληλοαναιρούνται .

$$f1_score = 2 * (precision * recall) / (precision + recall)$$

Στην συνέχεια, σε επίπεδο κώδικα για την αξιολόγηση της μηχανής αναζήτησης είναι αναγκαία η χρήση βιβλιοθηκών της Python για την υλοποίηση της αξιολόγησης. Μια τέτοια βιβλιοθήκη είναι η `pytrec_eval` . Για την εγκατάσταση της θα χρησιμοποιήσουμε την εντολή `-pip install pytrec_eval` στο terminal και για την χρήση της στον κώδικα την `import pytrec_eval`.

Συμπερασματικά, η επιλογή σωστής μετρικής εξαρτάται από τον τύπο του προβλήματος, τους στόχους της αξιολόγησης και ποια μορφή απόδοσης της μηχανής αναζήτησης θέλουμε να τονιστεί.

Αναφορά και Τεκμηρίωση (Βήμα 6°)

Κατά την υλοποίηση του προγράμματος ξεκινήσαμε με την δημιουργία του web crawler, το οποίο είναι μια συνάρτηση(**crawler**) που συλλέγει δεδομένα(τίτλος, συγγραφείς, abstract και ημερομηνία) από το arXiv με τη βοήθεια του BeautifulSoup και τα αποθηκεύει σε ένα .json αρχείο (**arXiv_results_raw.json**).

Στην συνέχεια, προχωρήσαμε στην προεπεξεργασία των δεδομένων που συλλέχθηκαν από τον web crawler(συνάρτηση **preprocess_text**). Πιο αναλυτικά, έγινε η αφαίρεση ειδικών χαρακτήρων, το tokenization, η κανονικοποίηση κειμένου, το Stemming, η αφαίρεση σημείων στίξης και προθημάτων (stop words) και η ενσωμάτωση των λέξεων σε ένα string. Τέλος, αυτά τα προεπεξεργασμένα δεδομένα καταχωρούνται στο arXiv_results_preprocessed.json αρχείο.

Έπειτα, έγινε η υλοποίηση του αντεστραμμένου ευρετηρίου(συνάρτηση **create_inverted_index**). Αρχικά η συνάρτηση συλλέγει τα έγγραφα και μετατρέπει το καθένα απ' αυτά σε λίστα στοιχείων. Μετά από κάποιους ελέγχους εντάσσονται στο ευρετήριο (**inverted_index.json**) και αυτό ταξινομείται αλφαβητικά.

Για να προχωρήσουμε στο επόμενο βήμα, πραγματοποιήθηκε μια εκτενής έρευνα αλγορίθμων και πληροφοριών έτσι ώστε να υλοποιηθούν: η Μηχανή Αναζήτησης, η Επεξεργασία ερωτήματος και η Κατάταξη αποτελεσμάτων. Αρχικά, αναπτύχθηκε μια γραφική διεπαφή χρήστη (συνάρτηση **user_interface**) στην οποία οι χρήστες έχουν την δυνατότητα να ανακτούν τα έγγραφα με τρεις διαφορετικούς αλγορίθμους. Για τη διεπαφή χρησιμοποιήσαμε τη βιβλιοθήκη της Python tkinter. Η διεπαφή αποτελείται από:

1. Πεδίο στο οποίο ο χρήστης εισάγει το ερώτημα του
2. 3 Radio buttons, ώστε να επιλέξει τον αλγόριθμο ανάκτησης της αρεσκείας του
3. 2 φίλτρα αναζήτησης. Τα φίλτρα δεν είναι υποχρεωτικά, δηλαδή κατά την αναζήτηση ακαδημαϊκών εργασιών γίνεται να εφαρμοστούν 1,2 ή και κανένα φίλτρο.
4. Κουμπί «Αναζήτηση», ώστε να ξεκινήσει η αναζήτηση(καλεί την εσωτερική συνάρτηση **search_and_display_results**).

Η εσωτερική συνάρτηση **search_and_display_results**, ασχολείται με την Επεξεργασία του Ερωτήματος του χρήστη. Αναλυτικότερα, λαμβάνει τις επιλογές του χρήστη στο GUI, και επιτελεί boolean πράξεις, στην περίπτωση που το ερώτημα αποτελείται από πάνω από 2 λέξεις. Παράλληλα, οφείλει να τονιστεί το γεγονός ότι για τις παραπάνω πράξεις δημιουργήσαμε τις βοηθητικές συναρτήσεις intersection, union, NOT. Τέλος, η εσωτερική συνάρτηση **search_and_display_results** καλεί τη συνάρτηση **engine** με τα κατάλληλα ορίσματα, η οποία αποτελεί τη σημαντικότερη συνάρτηση του προγράμματος. Αναλυτικότερα, η **engine** καλεί αρχικά τους 3 αλγορίθμους ανάκτησης που δημιουργήσαμε(ανάλογα με την επιλογή του χρήστη στο GUI), οι οποίοι είναι ο Boolean retrieval, ο VSM (Vector Space Model) και ο Okapi BM25.συναρτήσεις **boolean_retrieval**, **vector_space_model** και **okapi_bm25** αντίστοιχα). Ειδικότερα, ο αλγόριθμος boolean ανάκτησης υλοποιήθηκε με βάση τη Θεωρία του μαθήματος. Επιτελεί τις βασικές boolean πράξεις. Από την άλλη πλευρά, ο VSM και ο Okapi BM25 δημιουργήθηκαν μέσω αναζήτησης σε διάφορες πηγές στο διαδίκτυο. Όσον αφορά τον VSM, αρχικά κάνουμε προεπεξεργασία του ερωτήματος του χρήστη. Έπειτα, υλοποιούμε τον αλγόριθμο κατάταξης TF-IDF για το ερώτημα(έννοιες που έχουμε δει στη Θεωρία). Στη συνέχεια, με στόχο να ταξινομήσουμε τα έγγραφα ως προς την

ομοιότητα(από τον ορισμό του VSM), υπολογίζουμε την ομοιότητα συνημίτονου μεταξύ του ερωτήματος και των εγγράφων. Τέλος, για τον αλγόριθμο Okapi BM25, αν θέλαμε να χρησιμοποιήσουμε ορισμό, είναι μια συνάρτηση κατάταξης που χρησιμοποιείται από τις μηχανές αναζήτησης για την εκτίμηση της συνάφειας των εγγράφων με ένα δεδομένο ερώτημα αναζήτησης. Ακολουθεί το μοντέλο bag-of-words και ως τυπικές τιμές των παραμέτρων του χρησιμοποιούμε τις εξής: $k_1=1.5$ και $b=0.75$. Υπολογίζει το score αναζήτησης με βάση τη συνεισφορά κάθε όρου, λαμβάνοντας υπόψη τη συχνότητα εμφάνισης του όρου στο έγγραφο και το ερώτημα, καθώς και το μήκος του εγγράφου. Τα έγγραφα ταξινομούνται με βάση το score.

Σε αυτό το σημείο να υπενθυμίσουμε ότι συνεχίζεται η ανάλυση της συνάρτησης **engine**. Ανάλογα την επιλογή του χρήστη στο GUI επιλέγεται ο κατάλληλος αλγόριθμος και δημιουργούνται τα πρώτα αποτελέσματα(search_results). Στη συνέχεια, ανακτούμε τα συναφή με την αναζήτηση έγγραφα, με χρήση του ανεστραμμένου ευρετηρίου, που αναλύσαμε παραπάνω. Παράλληλα, η **engine** καλεί άλλες 3 συναρτήσεις:

1. **rank_documents**: Αρχικά γίνεται προεπεξεργασία του ερωτήματος. Στη συνέχεια συγχωνεύουμε τα συναφή έγγραφα σε μία συλλογή. Δημιουργήσαμε 2 τρόπους κατάταξης των αποτελεσμάτων. Πρώτα, υπολογίζουμε το TF-IDF(βασικός αλγόριθμος κατάταξης) για το συγκεκριμένο ερώτημα. Στη συνέχεια ταξινομούμε τα έγγραφα ως προς την ομοιότητα συνημίτονου μεταξύ του ερωτήματος και των εγγράφων(πιο προηγμένη τεχνική κατάταξης).
2. **filter_results**: Σε αυτή την συνάρτηση, έχοντας ως όρισμα τα αποτελέσματα της ανάκτησης, υλοποιούμε(εάν το επιλέξει ο χρήστης στο GUI) φίλτρα με βάση την ημερομηνία δημοσίευσης και το συγγραφέα του άρθρου που αναζητεί ο χρήστης. Γίνεται δηλαδή μία επιπλέον αναζήτηση στα αποτελέσματα, με σκοπό να βρεθούν συγκεκριμένα άρθρα, που συμβαδίζουν με τα παραπάνω φίλτρα.
3. **print_results**: Τέλος, η συγκεκριμένη συνάρτηση, έχει ως εργασία να εκτυπώσει τα τελικά αποτελέσματα στην οθόνη του χρήστη, σε φιλική μορφή(χρήση f-strings).

Σημειώσεις:

- 1) Το url της σελίδας που θα γίνει το crawling είναι το εξής:
<https://arxiv.org/list/cs/new>
- 2) Ο αριθμός των ακαδημαϊκών εργασιών που χρησιμοποιούμε στο σύνολο του προγράμματος είναι 150, αλλά ο κώδικας έχει δημιουργηθεί με δυναμικό τρόπο, ώστε να λειτουργεί για οποιοδήποτε αριθμό εργασιών.

Εν κατακλείδι, θα αναφέρουμε τις συναρτήσεις που καλούνται στη συνάρτηση `__main__`:

- **crawler**
- **save_to_json**(Αποθήκευση αρχείου JSON ΠΡΙΝ από το preprocess)
- **preprocess_text**
- **save_to_json**(Αποθήκευση αρχείου JSON META από το preprocess)
- **create_inverted_index**
- **save_to_json**(Αποθήκευση του ανεστραμμένου ευρετηρίου)
- **user_interface**(Καλεί όλες τις υπόλοιπες συναρτήσεις που εξηγήσαμε αναλυτικά παραπάνω)

Δυσκολίες - Βελτιώσεις

Παρ' όλα αυτά, κατά την εκπόνηση της εργασίας συναντήσαμε αρκετές δυσκολίες. Αναλυτικότερα, για την διαδικασία του web crawling , αρχικά επιλέξαμε τον ισότοπο Pub Med αλλά αντιμετωπίσαμε αδυναμία στην συλλογή των μεταδεδομένων της, συνεπώς, καταλήξαμε στη χρήση του arXiv στο οποίο συναντήσαμε πρόβλημα με τη συλλογή της ημερομηνίας(date) για κάθε έγγραφο. Στην συνέχεια, η εύρεση των αλγορίθμων ανάκτησης ήταν αρκετά περίπλοκη και περιορισμένη καθώς δεν υπήρχαν αρκετές πηγές προς μελέτη, γεγονός που μας οδηγεί στη διατήρηση επιφυλάξεων για την ορθή λειτουργία του VMS αλγορίθμου, αλλά και του Okapi BM25. Τελευταίο, αλλά εξίσου σημαντικό, στην διεπαφή υπάρχει αδυναμία στην εκτύπωση των τελικών αποτελεσμάτων. Αντιμετωπίσαμε αυτά τα προβλήματα πιθανώς λόγω περιορισμένου χρόνου. Ως ομάδα πιστεύουμε ότι ίσως μπορούσαν να βελτιωθούν σε κάποιο βαθμό, ωστόσο αξίζει να σημειωθεί ότι ήταν στόχος μας να υλοποιήσουμε όλα τα βήματα της εργασίας όσο καλύτερα μπορούσαμε.

Ακολουθούν φωτογραφίες που επιδεικνύουν την λειτουργικότητα του προγράμματος:

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\marin\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

- arXiv_results_preprocessed.json
- arXiv_results_raw.json
- inverted_index.json

arXiv_results_raw.json

```
{
  "arXiv_results_raw.json": [
    {
      "title": "Title:Transduce: learning transduction grammars for string transformation",
      "authors": [
        "Francis Frydman",
        "Philippe Mangion"
      ],
      "abstract": "The synthesis of string transformation programs from input-output examples\nutilizes various techniques, all based on an inductive bias that compr",
      "date": "Not Found"
    },
    {
      "title": "Title:RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models",
      "authors": [
        "Meiling Tao",
        "Xuechen Liang",
        "Tianyu Shi",
        "Lei Yu",
        "Yiting Xie"
      ],
      "abstract": "This study presents RoleCraft-GLM, an innovative framework aimed at enhancing\npersonalized role-playing with Large Language Models (LLMs). RoleCr",
      "date": "Not Found"
    },
    {
      "title": "Title:Modeling, Simulation, and Maneuvering Control of a Generic Submarine",
      "authors": [
        "Gage MacLin",
        "Maxwell Hammond",
        "Venzazio Cichella",
        "J. Ezequiel Martin"
      ],
      "abstract": "thi work introduc two multilevel control strategi to address the problem of guidanc and control of underwat vehicl an outerloop pathfollow algorith",
      "date": "Not Found"
    },
    {
      "title": "Title:Voxceleb-ESP: preliminary experiments detecting Spanish celebrities from their voices",
      "authors": [
        "Beltr\u00e9n Labrador",
        "Manuel Otero-Gonzalez",
        "Alicia Lozano-Diez",
        "Daniel Ramos",
        "Doroteo T. Toledano",
        "Joaquin Gonzalez-Rodriguez"
      ],
      "abstract": "thi paper present voxcelebesp a collect of pointer and timestamp to youtube video facilit the creation of a novel speaker recognit dataset voxcelebe",
      "date": "Not Found"
    },
    {
      "title": "Title:Object Attribute Matters in Visual Question Answering",
      "authors": [
        "Peize Li",
        "Qingyi Si",
        "Peng Fu",
        "Zheng Lin",
        "Yan Wang"
      ],
      "abstract": "visual question answer is a multimod task that requir the joint comprehens of visual and textual inform howev integr visual and textual semant sole",
      "date": "Not Found"
    }
  ]
}
```

arXiv_results_preprocessed.json


```
{
  "arXiv_results_preprocessed.json": [
    {
      "title": "Title:Transduce: learning transduction grammars for string transformation",
      "authors": [
        "Francis Frydman",
        "Philippe Mangion"
      ],
      "abstract": "The synthesis of string transformation programs from input-output examples\nutilizes various techniques, all based on an inductive bias that compr",
      "date": "Not Found"
    },
    {
      "title": "Title:RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models",
      "authors": [
        "Meiling Tao",
        "Xuechen Liang",
        "Tianyu Shi",
        "Lei Yu",
        "Yiting Xie"
      ],
      "abstract": "This study presents RoleCraft-GLM, an innovative framework aimed at enhancing\npersonalized role-playing with Large Language Models (LLMs). RoleCr",
      "date": "Not Found"
    },
    {
      "title": "Title:Modeling, Simulation, and Maneuvering Control of a Generic Submarine",
      "authors": [
        "Gage MacLin",
        "Maxwell Hammond",
        "Venzazio Cichella",
        "J. Ezequiel Martin"
      ],
      "abstract": "thi work introduc two multilevel control strategi to address the problem of guidanc and control of underwat vehicl an outerloop pathfollow algorith",
      "date": "Not Found"
    },
    {
      "title": "Title:Voxceleb-ESP: preliminary experiments detecting Spanish celebrities from their voices",
      "authors": [
        "Beltr\u00e9n Labrador",
        "Manuel Otero-Gonzalez",
        "Alicia Lozano-Diez",
        "Daniel Ramos",
        "Doroteo T. Toledano",
        "Joaquin Gonzalez-Rodriguez"
      ],
      "abstract": "thi paper present voxcelebesp a collect of pointer and timestamp to youtube video facilit the creation of a novel speaker recognit dataset voxcelebe",
      "date": "Not Found"
    },
    {
      "title": "Title:Object Attribute Matters in Visual Question Answering",
      "authors": [
        "Peize Li",
        "Qingyi Si",
        "Peng Fu",
        "Zheng Lin",
        "Yan Wang"
      ],
      "abstract": "visual question answer is a multimod task that requir the joint comprehens of visual and textual inform howev integr visual and textual semant sole",
      "date": "Not Found"
    }
  ]
}
```

inverted_index.json

```
{
  "inverted_index.json": {
    "abil": [
      37,
      70,
      6,
      136,
      105,
      77,
      144,
      28,
      95
    ],
    "abilitybecau": [
      38
    ],
    "abl": [
      50,
      95
    ],
    "ablat": [
      16
    ],
    "abnorm": [
      144
    ],
    "about": [
      137,
      75,
      11,
      143,
      84,
      150,
      59
    ],
    "abov": [
      70
    ],
    "abruptli": [
      125
    ],
    "absenc": [
      40
    ],
    "absent": [
      92
    ],
    "absolut": [
      25
    ]
  }
}
```

Διεπαφή χρήστη(GUI) για 2 διαφορετικά ερωτήματα χρηστών

Παράδειγμα 1(Απλό ερώτημα)

 Διεπαφή χρήστη για την αναζήτηση ακαδημαϊκών εργασιών

Ερώτημα χρήστη

access

Επιλογή αλγόριθμου ανάκτησης

☒ Boolean retrieval

☐ Vector Space Model(VSM)

☐ Probabilistic retrieval model(Okapi BM25)

Φίλτρα αναζήτησης

☐ Ημερομηνία δημοσίευσης

☐ Συγγραφέας

Αναζήτηση

Παράδειγμα 2(Επίδειξη λειτουργίας φίλτρων)

Διεπαφή χρήστη για την αναζήτηση ακαδημαϊκών εργασιών

Ερώτημα χρήστη

access AND accept

Επιλογή αλγόριθμου ανάκτησης

☐ Boolean retrieval

☒ Vector Space Model(VSM)

☐ Probabilistic retrieval model(Okapi BM25)

Φίλτρα αναζήτησης

☐ Ημερομηνία δημοσίευσης

☒ Συγγραφέας

Francesco Pasetti

Αναζήτηση

Αποτελέσματα που προκύπτουν από το 2^ο ερώτημα χρήστη

```

> doc_num = {list: 150} [{'abstract': 'the synthesi of string transform program from inputoutput exampl util variou techniqu all base on an induct bia that compris ...str transdu ... View
documents = {set: 1} {55}
  10
  01 {int} 55
  10
  01 __len__ = {int} 1
> Protected Attributes
  10
  01 i = {int} 150
  10
  01 inverted_index = {dict: 3227} {'abil': [37, 70, 6, 136, 105, 77, 144, 28, 95], 'abilitybecau': [38], 'abl': [50, 95], 'ablat': [16], 'abnorm': [144], 'about': [137, 75, 11, 143, 84, 150, 5 ... View
  10
  01 preprocessed_abstract = {str} 'onlin content platform commonli use engagementbas optim when make recommend thi encourag content creator to invest in qualiti but also... View
> res = {dict: 4} {'abstract': 'onlin content platform commonli use engagementbas optim when make recommend thi encourag content creator to invest in qualiti...aliz engag r... View
> results = {list: 580} [{'abstract': 'the synthesi of string transform program from inputoutput exampl util variou techniqu all base on an induct bia that compris ...str transduc l... View
  10
  01 term = {str} 'zt'
  10
  01 url = {str} 'https://arxiv.org/list/cs/new'
> Special Variables

```