Πανεπιστήμιο Ιωαννίνων, Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Ανάκτηση Πληροφορίας

Ακαδημαϊκό Έτος 2023-2024

Εργασία:

Μηχανή αναζήτησης πληροφορίας από επιστημονικά άρθρα

Ομάδα:

Αγγελική Γκαβαρδίνα, Α.Μ.: 4042

Αλέξανδρος Κόκκινος, Α.Μ.: 4084

Εισαγωγή

Το σύστημα που υλοποιήσαμε έχει στόχο να παρέχει μια ολοκληρωμένη λύση για τη συλλογή, ανάλυση και αναζήτηση επιστημονικών άρθρων από ένα CSV αρχείο. Η λειτουργικότητα του συστήματος περιλαμβάνει τη μετατροπή των δεδομένων του CSV αρχείου σε ευρετήριο που μπορεί να αναζητηθεί χρησιμοποιώντας τη βιβλιοθήκη Lucene, την υποστήριξη διαφόρων τύπων αναζητήσεων, και την παρουσίαση των αποτελεσμάτων αναζήτησης σε έναν γραφικό περιβάλλον χρήστη (GUI).

Συλλογή (corpus)

Χρησιμοποιήσαμε επιστημονικά άρθρα από την παρακάτω συλλογή από το kaggle: https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated/data?select=papers.csv . Η συλλογή περιλαμβάνει 200 επιστημονικά άρθρα, τα οποία πήραμε τυχαία μέσα από το αρχείο papers.csv . Τα πεδία του αρχείου αυτού είναι:

- **source_id:** μοναδικό id του επιστημονικού άρθρου
- **year:** χρονιά δημοσίευσης του επιστημονικού άρθρου
- **title:** τίτλος του επιστημονικού άρθρου
- **abstract:** περίληψη του επιστημονικού άρθρου
- **full_text:** το κείμενο του επιστημονικού άρθρου

Ανάλυση κειμένου και κατασκευή ευρετηρίου

Η διαδικασία που ακολουθήσαμε περιλαμβάνει τα εξής βήματα:

1. **Συλλογή Δεδομένων**:
    - Τα δεδομένα συλλέγονται από ένα CSV αρχείο (papers.csv) που περιέχει τις πληροφορίες source_id, year, title, abstract, και full_text.
    - Τα δεδομένα φορτώνονται στη μνήμη και αποθηκεύονται σε ένα HashSet από αντικείμενα τύπου Record.
2. **Προεπεξεργασία Άρθρων**:
    - Τα άρθρα μεταφέρονται σε ένα text αρχείο όπου κάθε εγγραφή περιλαμβάνει τα πεδία του CSV και διαχωρίζεται από τις άλλες με γραμμές ----------.
3. **Δημιουργία Ευρετηρίου**:
    - Χρησιμοποιείται η βιβλιοθήκη Apache Lucene για τη δημιουργία ευρετηρίου.
    - Τα πεδία που ευρετηριάζονται είναι τα source_id, year, title, abstract, και full_text.
    - Το πεδίο year αποθηκεύεται ως SortedDocValuesField για να υποστηρίζει αναζητήσεις με ταξινόμηση κατά έτος.
    - Χρησιμοποιείται EnglishAnalyzer για την ανάλυση του κειμένου, καθώς τα επιστημονικά άρθρα που χρησιμοποιούνται είναι στα αγγλικά.

Η υλοποίηση περιλαμβάνει τις ακόλουθες κλάσεις και μεθόδους:

- **DataCollector**: Διαχειρίζεται τη συλλογή και αποθήκευση των δεδομένων από το CSV αρχείο.
    - Μέθοδος **processCSV(String textFilePath)**: Διαβάζει τα δεδομένα από το CSV αρχείο και τα γράφει σε ένα text αρχείο.
    - Μέθοδος **getRecords()**: Επιστρέφει τα αρχεία που έχουν συλλεχθεί.
- **Index**: Δημιουργεί το ευρετήριο από τα δεδομένα.
    - Μέθοδος **indexTextFile(String textFilePath)**: Διαβάζει το text αρχείο και δημιουργεί το Lucene ευρετήριο.
    - Μέθοδος **close()**: Κλείνει τον IndexWriter.

Αναζήτηση

Το σύστημά μας υποστηρίζει τρεις κύριους τύπους αναζητήσεων:

1. **Αναζήτηση με λέξεις κλειδιά**:
    - Αναζητά λέξεις ή φράσεις μέσα στο πεδίο full_text χρησιμοποιώντας τον QueryParser της Lucene.
2. **Αναζήτηση σε συγκεκριμένα πεδία**:
    - Επιτρέπει αναζητήσεις σε συγκεκριμένα πεδία (title, abstract, year ή full_text).
    - Ο χρήστης μπορεί να καθορίσει το πεδίο και την τιμή που αναζητά, π.χ., title: machine learning.
3. **Αναζήτηση φράσεων**:

    - Αναζητά ακριβείς φράσεις μέσα στο full_text χρησιμοποιώντας το PhraseQuery της Lucene.

Η κλάση **Search** διαχειρίζεται τη διαδικασία αναζήτησης:

- **Search**: Υλοποιεί τις διάφορες μεθόδους αναζήτησης και διατηρεί ιστορικό αναζητήσεων.
    - Μέθοδος **keywordSearch(String queryStr)**: Αναζητά λέξεις κλειδιά.
    - Μέθοδος **fieldSearch(String field, String queryStr)**: Αναζητά σε συγκεκριμένο πεδίο.
    - Μέθοδος **phraseSearch(String queryStr)**: Αναζητά φράσεις.
    - Μέθοδος **getSearchHistory()**: Επιστρέφει το ιστορικό αναζητήσεων.

Παρουσίαση Αποτελεσμάτων

Τα αποτελέσματα της αναζήτησης επιλέξαμε να παρουσιάζονται σε έναν γραφικό περιβάλλον χρήστη (GUI) που υλοποιείται με χρήση του Swing. Η παρουσίαση περιλαμβάνει:

1. **Πίνακας Αποτελεσμάτων**:
    - Οι χρήστες μπορούν να δουν τα αποτελέσματα της αναζήτησης σε μορφή title, year, abstract. Η λέξη ή φράση που αναζητήθηκη γίνεται Highlight για να διευκολύνει το χρήστη. Τα αποτελέσματα εμφανίζονται ανά 10 σε σελίδες.
2. **Πλήρες Κείμενο**:
    - Η επιλογή ενός αποτελέσματος εμφανίζει το πλήρες κείμενο του άρθρου σε ένα νέο text παράθυρο.
3. **Δυνατότητες Αναδιάταξης**:
    - Τα αποτελέσματα μπορούν να αναδιαταχθούν βάση του πεδίου year.

Η κλάση **SearchGUI** διαχειρίζεται το γραφικό περιβάλλον χρήστη:

- **SearchGUI**: Υλοποιεί τη διασύνδεση και τις λειτουργικότητες αναζήτησης στο GUI.
    - Δέχεται την τοποθεσία του ευρετηρίου και τα δεδομένα Record.
    - Υλοποιεί τη διασύνδεση με το χρήστη για την εισαγωγή ερωτημάτων και την εμφάνιση των αποτελεσμάτων.

Υλοποίηση του Interface

Το γραφικό περιβάλλον χρήστη δημιουργήθηκε με τη χρήση της βιβλιοθήκης Swing. Οι βασικές λειτουργίες περιλαμβάνουν:

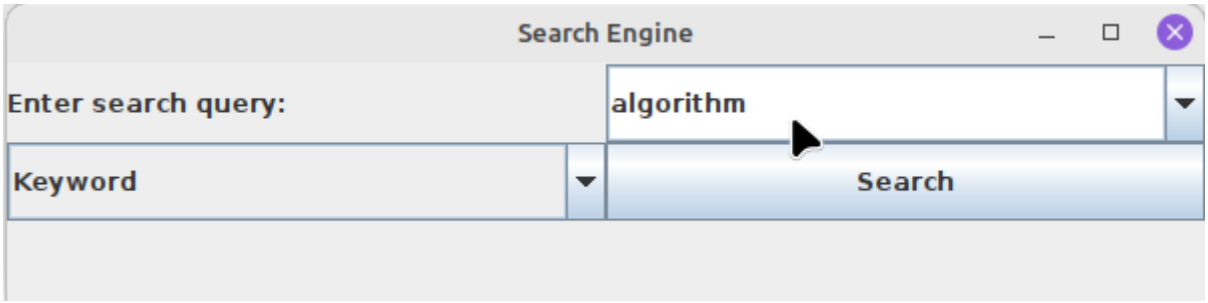
- **Πλαίσιο Εισαγωγής Αναζήτησης**:
    - Ένα πεδίο κειμένου για την εισαγωγή του ερωτήματος.
    - Μία αναπτυσσόμενη λίστα με τις 3 πιο πρόσφατες αναζητήσεις (ιστορικό).
    - Μια αναπτυσσόμενη λίστα για την επιλογή του τύπου αναζήτησης (Keyword, Field, Phrase).
    - Ένα κουμπί για την εκκίνηση της αναζήτησης.
- **Πίνακας Αποτελεσμάτων**:

    - Τα αποτελέσματα εμφανίζονται σε παράθυρο με τις σχετικές πληροφορίες.
- **Πεδίο Εμφάνισης Πλήρους Κειμένου**:
    - Το πλήρες κείμενο του επιλεγμένου άρθρου εμφανίζεται σε νέο παράθυρο.
- **Αναδιάταξη Αποτελεσμάτων**:
    - Οι χρήστες μπορούν να αναδιατάξουν τα αποτελέσματα βάση του πεδίου year.

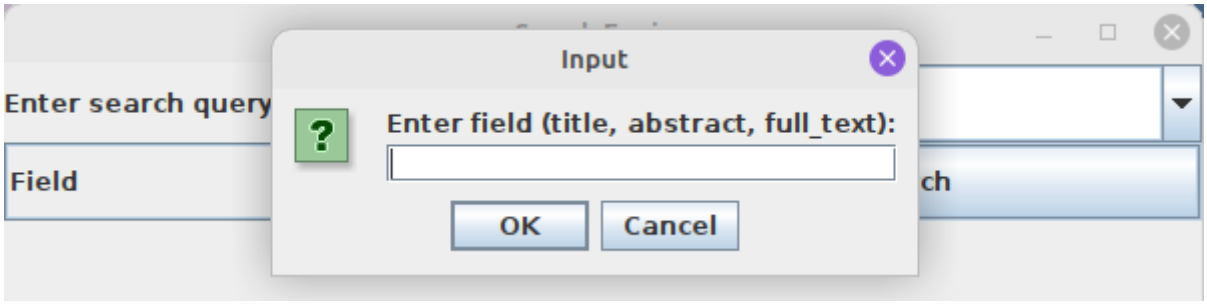
Η κλάση **SearchGUI** διαχειρίζεται όλες αυτές τις λειτουργίες, ενώ χρησιμοποιεί τις κλάσεις **Search** και **DataCollector** για να υλοποιήσει τη συνολική λειτουργικότητα του συστήματος.

Παράδειγμα Αναζήτησης

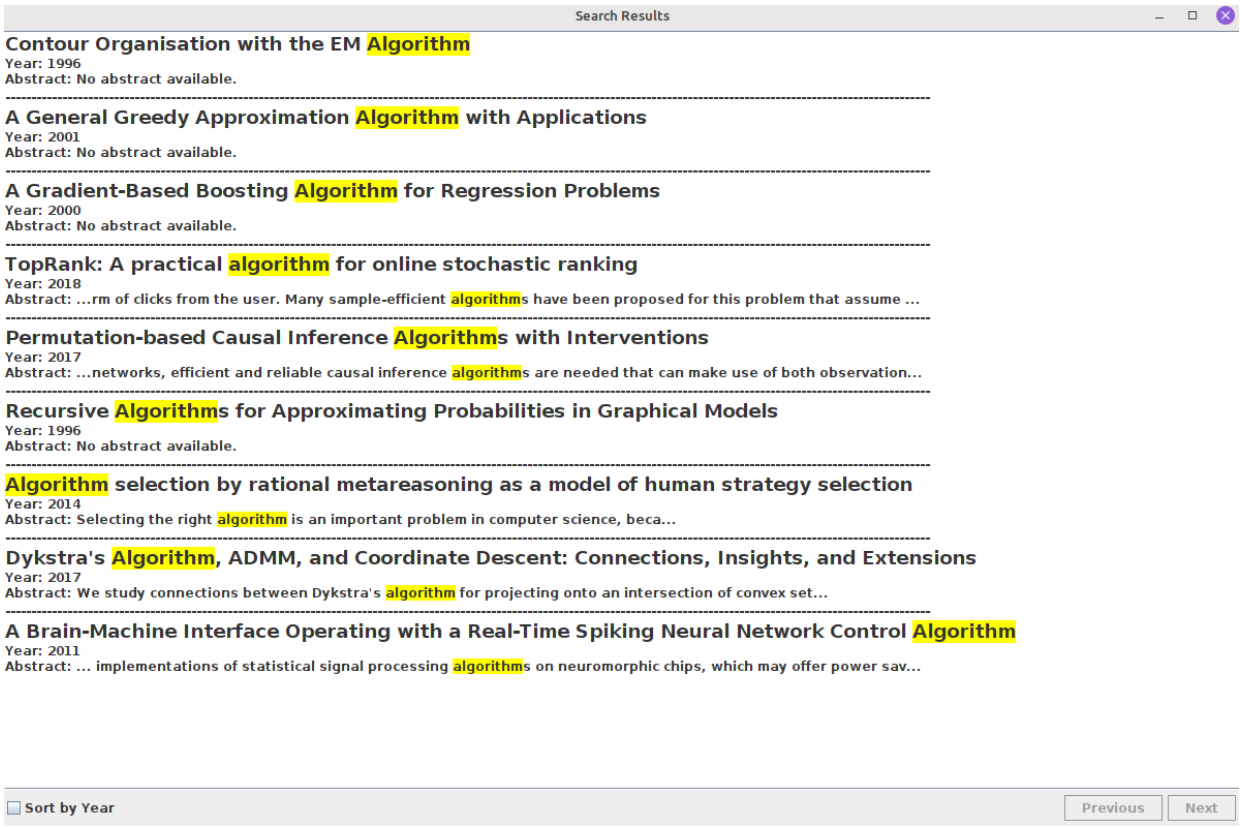
Ο χρήστης αναζητεί μια λέξη (π.χ. εδώ algorithm):



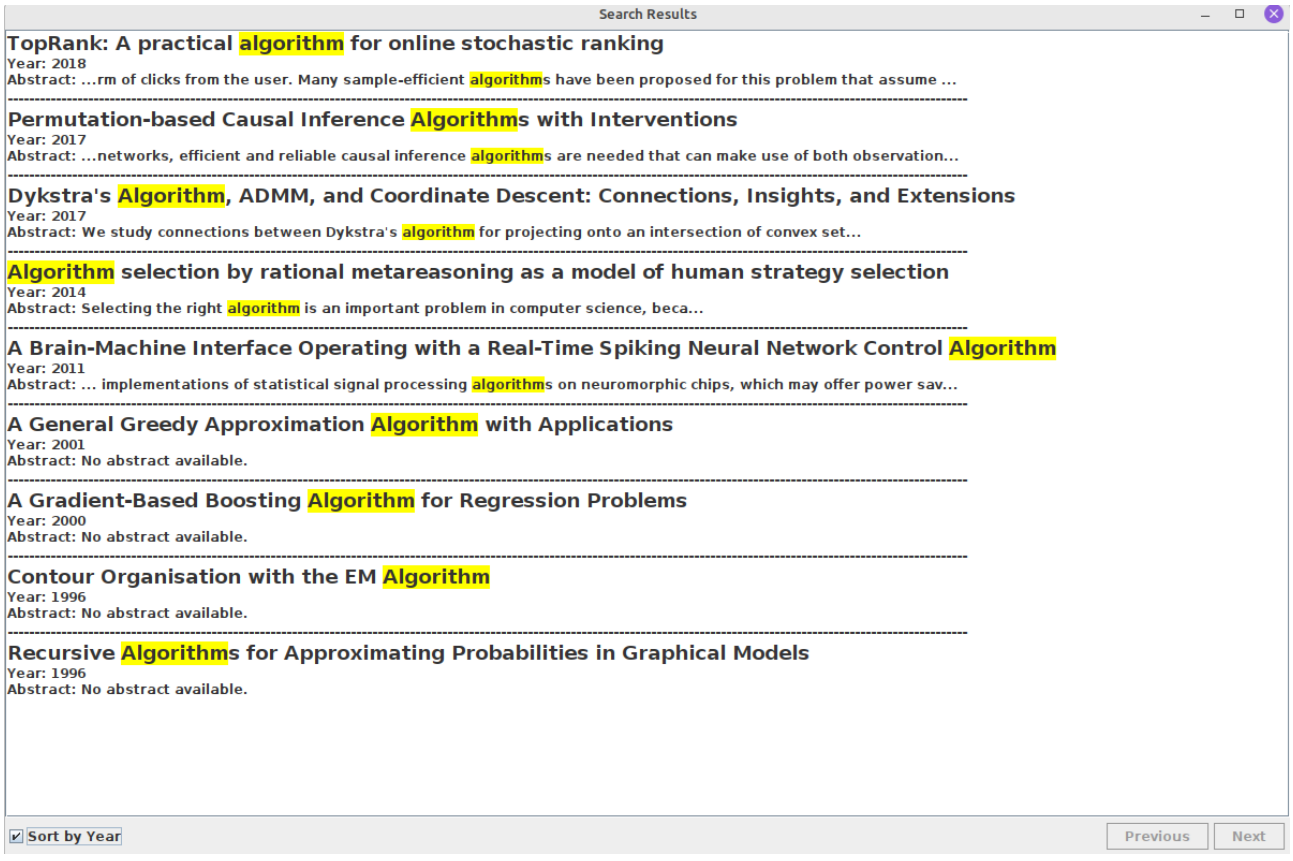
Επιλέγει αναζήτηση με πεδίο και πληκτρολογεί εκείνο που επιθυμεί (εδώ βάλαμε title):



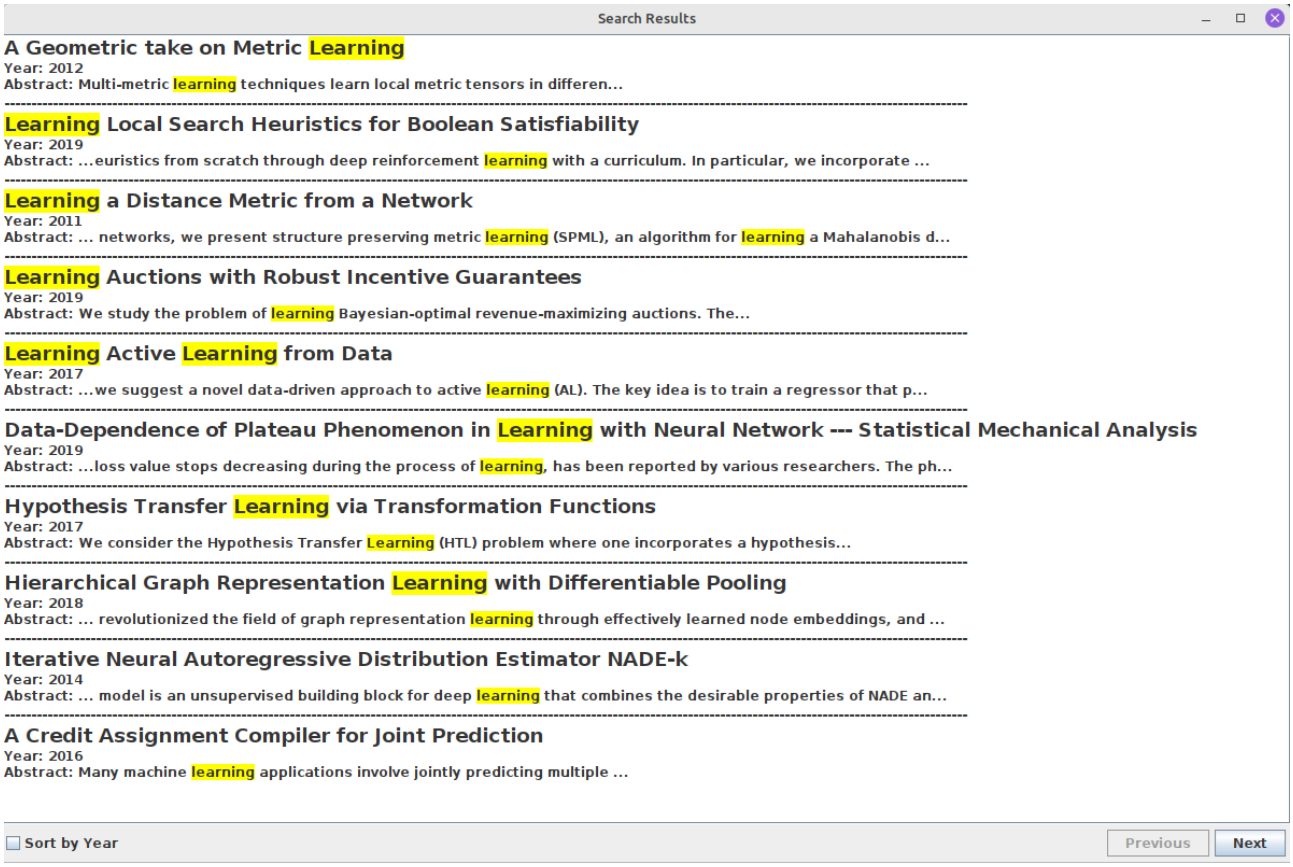
Επιλέγει OK και εμφανίζονται τα αποτελέσματα:



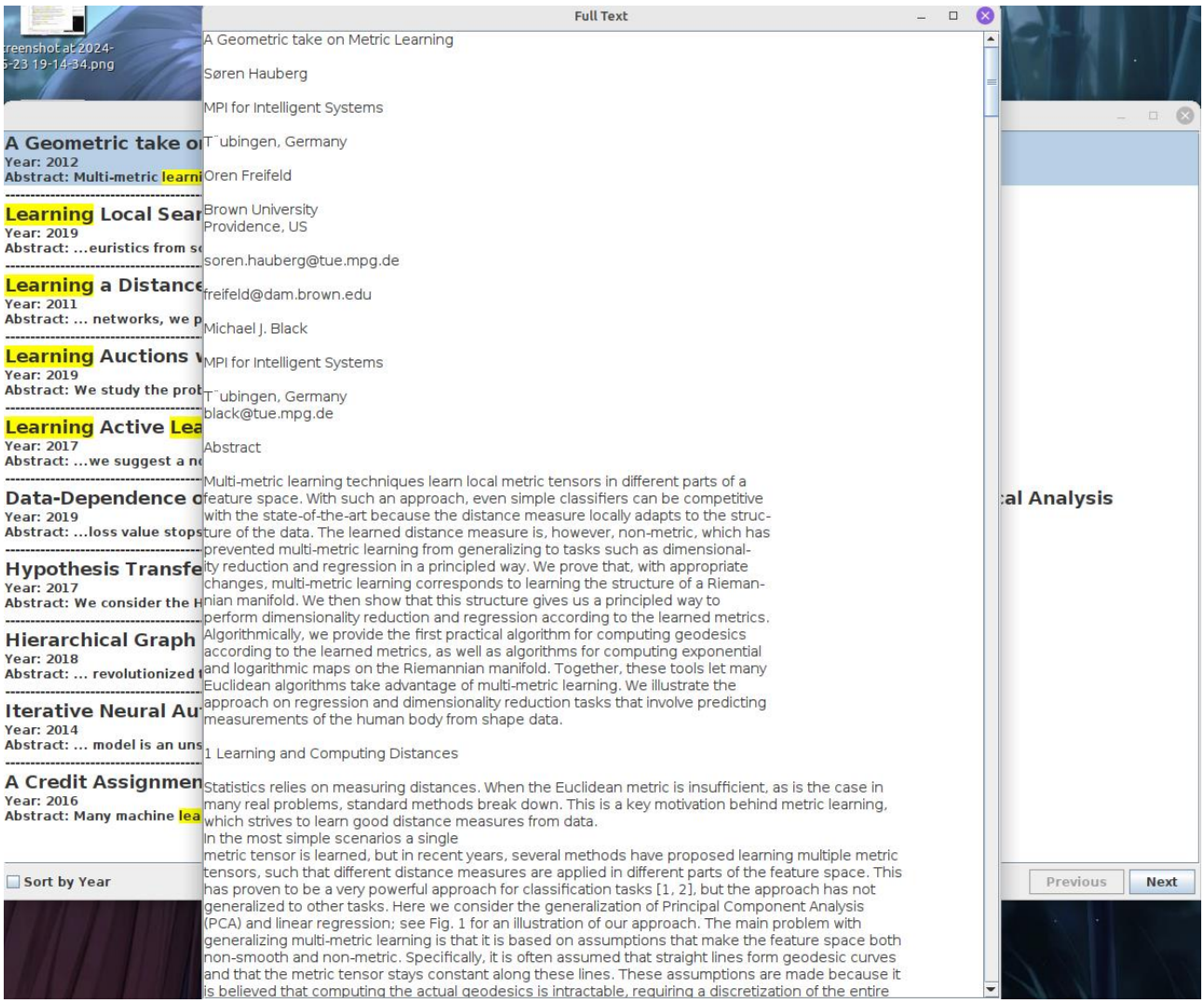
Αν επιλέξει το CheckBox Sort by Year τότε ταξινομούνται τα αποτελέσματα κατά χρονιά.



Ας αναζητήσουμε κάτι ακόμα. Αυτή τη φορά αναζητήσαμε τη λέξη Learning με πεδίο το abstract:



Αν πατήσω πάνω σε ένα αποτέλεσμα (π.χ. το πρώτο) μπορώ να διαβάσω ολόκληρο το κείμενο:



Ευχαριστούμε για το χρόνο σας.