

Εργασία:

Μηχανή αναζήτησης πληροφορίας από επιστημονικά άρθρα

Ομάδα:

Αγγελική Γκαβαρδίνη, 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/rowhitsuami/nips-papers-1987-2019-](https://www.kaggle.com/datasets/rowhitsuami/nips-papers-1987-2019-updated/data?select=papers.csv)

[updated/data?select=papers.csv](https://www.kaggle.com/datasets/rowhitsuami/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 υποστηρίξει 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 αναζητήθηκε 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.

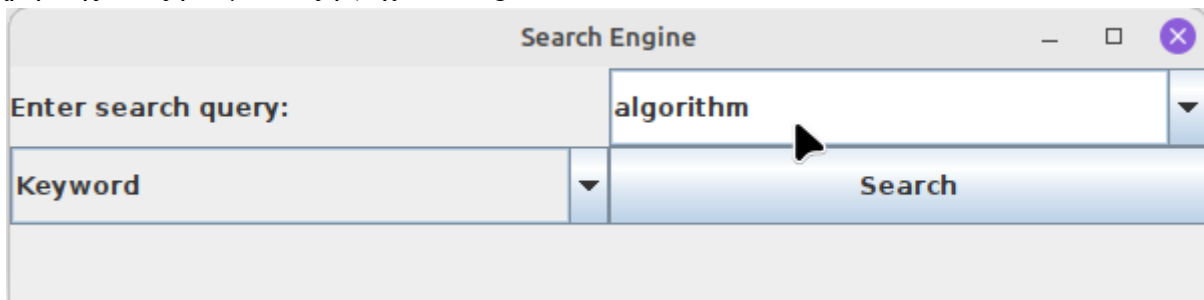
• Πίνακας Αποτελεσμάτων:

- 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.

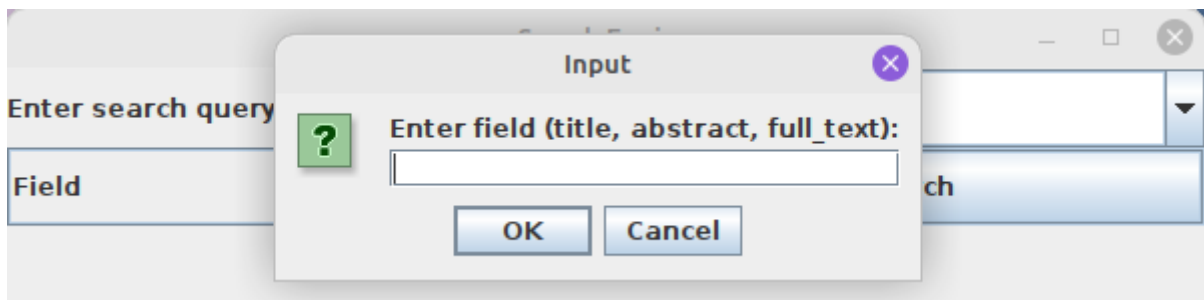
Η κλάση **SearchGUI** διαχειρίζεται όλες αυτές τις λειτουργίες, ενώ χρησιμοποιεί τις κλάσεις **Search** και **DataCollector** για να υλοποιήσει τη συνολική λειτουργικότητα του συστήματος.

Παράδειγμα Αναζήτησης

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



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



Επιλέγει OK και εμφανίζονται τα αποτελέσματα:

Search Results

Contour Organisation with the EM Algorithm

Year: 1996

Abstract: No abstract available.

A General Greedy Approximation Algorithm with Applications

Year: 2001

Abstract: No abstract available.

A Gradient-Based Boosting Algorithm for Regression Problems

Year: 2000

Abstract: No abstract available.

TopRank: A practical algorithm for online stochastic ranking

Year: 2018

Abstract: ...rm of clicks from the user. Many sample-efficient algorithms have been proposed for this problem that assume ...

Permutation-based Causal Inference Algorithms with Interventions

Year: 2017

Abstract: ...networks, efficient and reliable causal inference algorithms are needed that can make use of both observation...

Recursive Algorithms for Approximating Probabilities in Graphical Models

Year: 1996

Abstract: No abstract available.

Algorithm selection by rational metareasoning as a model of human strategy selection

Year: 2014

Abstract: Selecting the right algorithm is an important problem in computer science, beca...

Dijkstra's Algorithm, ADMM, and Coordinate Descent: Connections, Insights, and Extensions

Year: 2017

Abstract: We study connections between Dijkstra's algorithm for projecting onto an intersection of convex set...

A Brain-Machine Interface Operating with a Real-Time Spiking Neural Network Control Algorithm

Year: 2011

Abstract: ... implementations of statistical signal processing algorithms on neuromorphic chips, which may offer power sav...

☐ Sort by Year

Previous

Next

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

Search Results

TopRank: A practical algorithm for online stochastic ranking

Year: 2018

Abstract: ...rm of clicks from the user. Many sample-efficient algorithms have been proposed for this problem that assume ...

Permutation-based Causal Inference Algorithms with Interventions

Year: 2017

Abstract: ...networks, efficient and reliable causal inference algorithms are needed that can make use of both observation...

Dijkstra's Algorithm, ADMM, and Coordinate Descent: Connections, Insights, and Extensions

Year: 2017

Abstract: We study connections between Dijkstra's algorithm for projecting onto an intersection of convex set...

Algorithm selection by rational metareasoning as a model of human strategy selection

Year: 2014

Abstract: Selecting the right algorithm is an important problem in computer science, beca...

A Brain-Machine Interface Operating with a Real-Time Spiking Neural Network Control Algorithm

Year: 2011

Abstract: ... implementations of statistical signal processing algorithms on neuromorphic chips, which may offer power sav...

A General Greedy Approximation Algorithm with Applications

Year: 2001

Abstract: No abstract available.

A Gradient-Based Boosting Algorithm for Regression Problems

Year: 2000

Abstract: No abstract available.

Contour Organisation with the EM Algorithm

Year: 1996

Abstract: No abstract available.

Recursive Algorithms for Approximating Probabilities in Graphical Models

Year: 1996

Abstract: No abstract available.

☒ Sort by Year

Previous

Next

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

Search Results

A Geometric take on Metric Learning
Year: 2012
Abstract: Multi-metric learning techniques learn local metric tensors in differen...

Learning Local Search Heuristics for Boolean Satisfiability
Year: 2019
Abstract: ...euristics from scratch through deep reinforcement learning with a curriculum. In particular, we incorporate ...

Learning a Distance Metric from a Network
Year: 2011
Abstract: ... networks, we present structure preserving metric learning (SPML), an algorithm for learning a Mahalanobis d...

Learning Auctions with Robust Incentive Guarantees
Year: 2019
Abstract: We study the problem of learning Bayesian-optimal revenue-maximizing auctions. The...

Learning Active Learning from Data
Year: 2017
Abstract: ...we suggest a novel data-driven approach to active learning (AL). The key idea is to train a regressor that p...

Data-Dependence of Plateau Phenomenon in Learning with Neural Network --- Statistical Mechanical Analysis
Year: 2019
Abstract: ...loss value stops decreasing during the process of learning, has been reported by various researchers. The ph...

Hypothesis Transfer Learning via Transformation Functions
Year: 2017
Abstract: We consider the Hypothesis Transfer Learning (HTL) problem where one incorporates a hypothesis...

Hierarchical Graph Representation Learning with Differentiable Pooling
Year: 2018
Abstract: ... revolutionized the field of graph representation learning through effectively learned node embeddings, and ...

Iterative Neural Autoregressive Distribution Estimator NADE-k
Year: 2014
Abstract: ... model is an unsupervised building block for deep learning that combines the desirable properties of NADE an...

A Credit Assignment Compiler for Joint Prediction
Year: 2016
Abstract: Many machine learning applications involve jointly predicting multiple ...

☐ Sort by Year

PreviousNext

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

Full Text

Thumbnail

A Geometric take on Metric Learning

Søren Hauberg

MPI for Intelligent Systems

Tübingen, Germany

Year: 2012

Abstract: Multi-metric learning

Learning Local Search

Year: 2019

Abstract: ...heuristics from search

Learning a Distance

Year: 2011

Abstract: ... networks, we propose

Learning Auctions

Year: 2019

Abstract: We study the problem of

Learning Active Learning

Year: 2017

Abstract: ...we suggest a new

Data-Dependence of

Year: 2019

Abstract: ...loss value stops

Hypothesis Transfer

Year: 2017

Abstract: We consider the hypothesis

Hierarchical Graph

Year: 2018

Abstract: ... revolutionized the

Iterative Neural Auto

Year: 2014

Abstract: ... model is an unsupervised

A Credit Assignment

Year: 2016

Abstract: Many machine learning

Sort by Year

Thumbnail

A Geometric take on Metric Learning

Søren Hauberg

MPI for Intelligent Systems

Tübingen, Germany

Year: 2012

Abstract: Multi-metric learning techniques learn local metric tensors in different parts of a feature space. With such an approach, even simple classifiers can be competitive with the state-of-the-art because the distance measure locally adapts to the structure of the data. The learned distance measure is, however, non-metric, which has prevented multi-metric learning from generalizing to tasks such as dimensionality reduction and regression in a principled way. We prove that, with appropriate changes, multi-metric learning corresponds to learning the structure of a Riemannian manifold. We then show that this structure gives us a principled way to perform dimensionality reduction and regression according to the learned metrics. Algorithmically, we provide the first practical algorithm for computing geodesics according to the learned metrics, as well as algorithms for computing exponential and logarithmic maps on the Riemannian manifold. Together, these tools let many Euclidean algorithms take advantage of multi-metric learning. We illustrate the approach on regression and dimensionality reduction tasks that involve predicting measurements of the human body from shape data.

1 Learning and Computing Distances

Statistics relies on measuring distances. When the Euclidean metric is insufficient, as is the case in many real problems, standard methods break down. This is a key motivation behind metric learning, which strives to learn good distance measures from data. In the most simple scenarios a single metric tensor is learned, but in recent years, several methods have proposed learning multiple metric tensors, such that different distance measures are applied in different parts of the feature space. This has proven to be a very powerful approach for classification tasks [1, 2], but the approach has not generalized to other tasks. Here we consider the generalization of Principal Component Analysis (PCA) and linear regression; see Fig. 1 for an illustration of our approach. The main problem with generalizing multi-metric learning is that it is based on assumptions that make the feature space both non-smooth and non-metric. Specifically, it is often assumed that straight lines form geodesic curves and that the metric tensor stays constant along these lines. These assumptions are made because it is believed that computing the actual geodesics is intractable, requiring a discretization of the entire

Local Analysis

Previous

Next

Ευχαριστούμε για το χρόνο σας.