

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΙΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΔΗΜΙΟΥΡΓΙΑ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ
ΑΚΑΔΗΜΑΪΚΩΝ ΕΡΓΑΣΙΩΝ

ΣΤΟΙΧΕΙΑ ΕΡΓΑΣΙΑΣ

ΥΠΕΥΘΥΝΗ ΕΡΓΑΣΤΗΡΙΟΥ : ΤΣΕΛΕΝΤΗ ΠΑΝΑΓΙΩΤΑ

ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ : 29/1/2024

ΠΡΟΘΕΣΜΙΑ ΥΠΟΒΟΛΗΣ : 29/1/2024

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ 1

ΦΩΤΟΓΡΑΦΙΑ ΦΟΙΤΗΤΗ:



ΟΝΟΜΑΤΕΠΩΝΥΜΟ : ΑΘΑΝΑΣΙΟΥ ΒΑΣΙΛΕΙΟΣ ΕΥΑΓΓΕΛΟΣ

ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ : 19390005

ΕΞΑΜΗΝΟ ΦΟΙΤΗΤΗ : 9^ο

ΚΑΤΑΣΤΑΣΗ ΦΟΙΤΗΤΗ : ΠΡΟΠΤΥΧΙΑΚΟ

ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ : ΠΑΔΑ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ 2

ΦΩΤΟΓΡΑΦΙΑ ΦΟΙΤΗΤΗ:



ΟΝΟΜΑΤΕΠΩΝΥΜΟ : ΤΑΤΣΗΣ ΠΑΝΤΕΛΗΣ

ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ : 20390226

ΕΞΑΜΗΝΟ ΦΟΙΤΗΤΗ : 7^ο

ΚΑΤΑΣΤΑΣΗ ΦΟΙΤΗΤΗ : ΠΡΟΠΤΥΧΙΑΚΟ

ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ : ΠΑΔΑ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΠΕΡΙΕΧΟΜΕΝΑ

Βήματα	5
1. Σταχυολογητής (Web Crawler)	5
1.α. Επιλογή ιστότοπου-στόχου	5
1.β. Υλοποίηση web crawler	9
1.γ. Αποθήκευση δεδομένων σε δομημένη μορφή	14
2. Προεπεξεργασία κειμένου (Text Processing).....	16
2.α. Σχεδιασμός	16
2.β. Υλοποίηση.....	16
2.γ. Αξιολόγηση	17
2.δ. Εφαρμογή	18
3. Ευρετήριο (Indexing)	22
3.α. Δημιουργία της ανεστραμμένης δομής δεδομένων ευρετηρίου	22
3.β. Αποθήκευση του ευρετηρίου σε μία δομή δεδομένων.....	23
4. Μηχανή αναζήτησης (Search Engine).....	24
4.α. Ανάπτυξη διεπαφής χρήστη για αναζήτηση εργασιών.....	24
4.β. Υλοποίηση αλγορίθμων ανάκτησης.....	25
4.γ. Φιλτράρισμα αποτελεσμάτων αναζήτησης με διάφορα κριτήρια	40
Επεξεργασία ερωτήματος (Query Processing)	43
Κατάταξη αποτελεσμάτων (Ranking)	45
5. Αξιολόγηση συστήματος.....	49
5.α. Σύνολα Δεδομένων (Dataset)	49
5.β. Σενάρια Αξιολόγησης.....	49
5.γ. Βιβλιοθήκες Python.....	49
5.δ. Εφαρμογές μετρικών	49
5.ε. Ανάλυση και Βελτιώσεις	49

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Βήματα

1. Σταχνολογητής (Web Crawler)

1.a. Επιλογή ιστότοπου-στόχου

1.a.1 Σχεδιασμός

Ο ιστότοπος-στόχος είναι το αποθετήριο ακαδημαϊκών εργασιών [arXiv](#). Το σχεδιαστικό μοντέλο αφορά στην ανάκτηση της σελίδας με τα αποτελέσματα που παράγει η μηχανή αναζήτησης του ιστότοπου από ένα ή και περισσότερα τυχαία ερώτηματα χρήστη, με απότερο σκοπό την δημιουργία του αποθετηρίου (dataset) της τοπικής μηχανής αναζήτησης, πάνω στο οποίο θα γίνονται οι αναζητήσεις με ερωτήματα χρήστη.

The screenshot shows the arXiv homepage. At the top, there's a navigation bar with the Cornell University logo, a "We gratefully acknowledge support from the Simons Foundation member institutions, and all contributors. Donate" message, and a "Login" button. Below the header is the arXiv logo. A search bar with fields for "Search", "All fields", and a dropdown menu is present. To the right of the search bar are "Help" and "Advanced Search" links. A "arXiv News" section with a "Stay up to date with what is happening at arXiv on our blog." link is visible. A "Latest news" section follows. The main content area is divided into sections: "Physics" and "Mathematics". The "Physics" section includes a "Subject search and browse" dropdown set to "Physics", a "Search" button, and links for "Form Interface" and "Catchup". The "Physics" section lists various sub-fields like Astrophysics, Condensed Matter, General Relativity, High Energy Physics, Mathematical Physics, Nonlinear Sciences, Nuclear Theory, Physics, and Quantum Physics, each with a "new", "recent", and "search" link. The "Mathematics" section also lists sub-fields with similar links. At the bottom, a "Computer Science" section is partially visible.

Εικόνα 1.A.1 Ο ιστότοπος-στόχος arXiv

1.a.2 Υλοποίηση

Αρχικά, το πλήθος των ερωτημάτων χρήστη και η επιλογή αυτών είναι τυχαία σε κάθε εκτέλεση της τοπικής μηχανής αναζήτησης (main.py). Τα ερωτήματα περιορίζονται από 2 έως 8 και συγκεκριμένα έχει επιλεχθεί να είναι οι τίτλοι των βασικών μαθημάτων πάνω στα οποία έχουν αναρτηθεί ακαδημαϊκές εργασίες. Το μοντέλο υλοποίησης αφορά στην αποστολή ενός αιτήματος HTTP-GET στον [σύνδεσμο](#) της μηχανής αναζήτησης του ιστοτόπου μαζί με το αντίστοιχο τυχαίο ερώτημα χρήστη για την εμφάνιση αποτελεσμάτων. Η διαδικασία είναι επαναληπτική εφόσον, το πλήθος των ερωτημάτων είναι περισσότερο από 1. Αξίζει να σημειωθεί ότι για κάθε ερώτημα, έχει επιλεχθεί ο σύνδεσμος που εμφανίζει 100 αποτελέσματα ακαδημαϊκών

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

εργασιών. Για παράδειγμα, για τα ερωτήματα χρήστη Physics, Statistics και Computer, αποστέλλονται 3 HTTP-GET αιτήματα στους αντίστοιχους συνδέσμους:

https://arxiv.org/search/?query=Physics&searchtype=all&abstracts=show&order=-announced_date_first&size=100

https://arxiv.org/search/?query=Statistics&searchtype=all&abstracts=show&order=-announced_date_first&size=100

https://arxiv.org/search/?query=Computer&searchtype=all&abstracts=show&order=-announced_date_first&size=100

The screenshot shows the arXiv search interface with a red header. It features two main sections: 'Searching by Author Name' and 'Searching by subcategory'. Both sections provide tips for effective searching. Below these are 'Tips' and a section titled 'Εικόνα 1.A.2 Η μηχανή αναζήτησης του arXiv'.

Searching by Author Name

- Using the **Author(s)** field produces best results for author name searches.
- For the most precise name search, follow **surname(s), forename(s) or surname(s), initial(s)** pattern: example Hawking, S or Hawking, Stephen
- For best results on multiple author names, **separate individuals with a ;** (semicolon). Example: Jin, D S; Ye, J
- Author names enclosed in quotes will return only **exact matches**. For example, "Stephen Hawking" will not return matches for Stephen W. Hawking.
- Diacritic character variants are automatically searched in the Author(s) field.
- Queries with no punctuation will treat each term independently.

Searching by subcategory

- To search within a subcategory select **All fields**.
- A subcategory search can be combined with an author or keyword search by clicking on **add another term** in advanced search.

The screenshot shows the arXiv search results for 'all: Physics'. It displays three results, each with a title, authors, abstract, and download links. The results are paginated at the bottom.

Showing 1–100 of 919,883 results for all: Physics

1. arXiv:2401.14402 [pdf, other] [abs](#) [pdf](#) [HE](#)
Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors: M. Reichert, M. Bugli, J. Guillet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Abstract: ...ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear **physics** uncertainties such as different nuclear masses, β^- -decays and β^- -delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy... [More](#)
Submitted 25 January, 2024, originally announced January 2024.

2. arXiv:2401.14398 [pdf, other] [abs](#) [CV](#) [cs](#) [LG](#)
pix2gestalt: Amodal Segmentation by Synthesizing Wholes
Authors: Ege Ozguroglu, Ruoshi Liu, Didier Suris, Dian Chen, Achal Dave, Pavel Tokmakov, Carl Vondrick
Abstract: ...representations to this task, we learn a conditional diffusion model for reconstructing whole objects in challenging zero-shot cases, including examples that break natural and **physical** priors, such as art. As training data, we use a synthetically curated dataset containing occluded objects paired with their whole counterparts. Experiments show that our appro... [More](#)
Submitted 25 January, 2024, originally announced January 2024.
Comments: Website: <https://gestalt.cs.columbia.edu/>

3. arXiv:2401.14396 [pdf, other] [abs](#) [cond-mat.stat-mech](#) [hep-lat](#)
Entanglement entropy and deconfined criticality: emergent SO(5) symmetry and proper lattice bipartition

Εικόνα 1.A.3 Τα πρώτα 100 αποτελέσματα αναζήτησης του ερωτήματος Physics

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

The screenshot shows the arXiv search results for the query "Statistics". The page title is "Showing 1–100 of 221,264 results for all: Statistics". The search bar contains "Statistics". Below the search bar are buttons for "Show abstracts" (selected) and "Hide abstracts". To the right are links for "Search v0.5.6 released 2020-02-24", "Feedback?", "All fields", "Advanced Search", and "Login". A red banner at the top right reads "We gratefully acknowledge support from the Simons Foundation and member institutions.".

Showing 1–100 of 221,264 results for all: Statistics

Search v0.5.6 released 2020-02-24 Feedback?

All fields Advanced Search

Login

Statistics

Search

Show abstracts Hide abstracts

Advanced Search

100 results per page Sort results by Announcement date (newest first) Go

1 2 3 4 5 ... Next

1. arXiv:2401.14393 [pdf, other] stat.AP stat.ME

Clustering-based spatial interpolation of parametric post-processing models

Authors: Sándor Baran, Mária Lakatos

Abstract: ...development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of statistical post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated w... ▾ More

Submitted 25 January, 2024; originally announced January 2024.

Comments: 19 pages, 6 figures

2. arXiv:2401.14378 [pdf, ps, other] cond-mat.stat-mech

Single-file dynamics with general charge measures

Authors: Ziga Krajnik

Abstract: ...transformation acting on the finite-time distribution of particle fluctuations. The transformation is mapped to a simple substitution rule for corresponding full-counting statistics. By taking the asymptotics of the dressing transformation we analyze typical and large scale charge fluctuations.

Typical charge fluctuations in equilibrium states with vanishing... ▾ More

Submitted 25 January, 2024; originally announced January 2024.

Comments: 18+9 pages

3. arXiv:2401.14372 [pdf, other] physics.geo-ph

Εικόνα 1.Α.4 Τα πρώτα 100 αποτελέσματα αναζήτησης του ερωτήματος Statistics

The screenshot shows the arXiv search results for the query "Computer". The page title is "Showing 1–100 of 735,886 results for all: Computer". The search bar contains "Computer". Below the search bar are buttons for "Show abstracts" (selected) and "Hide abstracts". To the right are links for "Search v0.5.6 released 2020-02-24", "Feedback?", "All fields", "Advanced Search", and "Login". A red banner at the top right reads "We gratefully acknowledge support from the Simons Foundation and member institutions.".

Showing 1–100 of 735,886 results for all: Computer

Search v0.5.6 released 2020-02-24 Feedback?

All fields Advanced Search

Login

Computer

Search

Show abstracts Hide abstracts

Advanced Search

100 results per page Sort results by Announcement date (newest first) Go

1 2 3 4 5 ... Next

1. arXiv:2401.14405 [pdf, other] cs.CV cs.AI cs.LG

Multimodal Pathway: Improve Transformers with Irrelevant Data from Other Modalities

Authors: Yijuan Zhang, Xiaohan Ding, Kaixiong Gong, Yuxiao Ge, Ying Shan, Xiangyu Yue

Abstract: We propose to improve transformers of a specific modality with irrelevant data from other modalities, e.g., improve an ImageNet model with audio point cloud datasets. We would like to highlight that the data samples of the target modality are irrelevant to the other modalities, which distinguishes our method from other works utilizing paired (e.g., CLIP) or interleaved data of different modalit... ▾ More

Submitted 25 January, 2024; originally announced January 2024.

Comments: The code and models are available at <https://github.com/AI4Lab-CVCM2P>

2. arXiv:2401.14404 [pdf, other] cs.CV cs.I.G

Deconstructing Denoising Diffusion Models for Self-Supervised Learning

Authors: Xinlei Chen, Zhuang Liu, Saining Xie, Kaiming He

Abstract: In this study, we examine the representation learning abilities of Denoising Diffusion Models (DDM) that were originally purposed for image generation. Our philosophy is to deconstruct a DDM, gradually transforming it into a classical Denoising Autoencoder (DAE). This deconstructive procedure allows us to explore how various components of modern DDMs influence self-supervised representation learni... ▾ More

Submitted 25 January, 2024; originally announced January 2024.

Comments: Technical report, 10 pages

3. arXiv:2401.14403 [pdf, other] cs.RO cs.AI cs.CV cs.LG eess.SY

Εικόνα 1.Α.5 Τα πρώτα 100 αποτελέσματα αναζήτησης του ερωτήματος Computer

1.α.3 Αξιολόγηση

Η τοπική μηχανή υποστηρίζει χειρισμό εξαιρέσεων για την περίπτωση που ο κωδικός επιστροφής ενός αιτήματος HTTP-GET είναι διαφορετικός από 200. Διεξοδικά, ο κωδικός 200 σημαίνει ότι το αίτημα εξυπηρετήθηκε και η σελίδα πάνω στην οποία έγινε το αίτημα ανακτήθηκε με επιτυχία. Αξίζει να σημειωθεί ότι η επιλογή των τυχαίων ερωτημάτων χρήστη για την δημιουργία του dataset είναι από 2 έως 8, ώστε το μέγεθος του dataset να είναι από 200 έως 800 εργασίες.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

1.β. Υλοποίηση web crawler

1.β.1 Σχεδιασμός

Ο σταχυολογητής (web crawler) έχει σχεδιαστεί προκειμένου, να συλλέγει τα μεταδεδομένα των ακαδημαϊκών εργασιών (web scrape) από την σελίδα που εμφάνισε η μηχανή αναζήτησης του ιστότοπου με είσοδο κάποιο από τα τυχαία ερωτήματα χρήστη. Η διαδικασία είναι επαναληπτική από το γεγονός που τεκμηριώθηκε στην [υλοποίηση](#) του βήματος 1.α Επιλογή ιστότοπου-στόχου.

1.β.2 Υλοποίηση

Η υλοποίηση ακολουθεί τον web-crawler BeautifulSoup της γλώσσας Python και λαμβάνει χώρα σε ξεχωριστό module (web_crawler.py), όπου επιστρέφει την σελίδα που αιτήθηκε για ανάκτηση μέσω ενός επιτυχημένου αιτήματος HTTP-GET, σε μη-δομημένη μορφή HTML. Τα δεδομένα που συλλέγονται από τις εργασίες είναι τα εξής:

- Τίτλος
- Συγγραφείς
- Μαθήματα και υπο-μαθήματα που σχετίζονται
- Περίληψη
- Σχόλια
- Ημερομηνία δημοσίευσης
- Σύνδεσμος για την λήψη της εργασίας σε μορφή pdf

Στην υλοποίηση προστίθεται στα μεταδεδομένα της κάθε εργασίας και ένας αναγνωριστικός ακέραιος αριθμός (doc_id), ο οποίος είναι μοναδικός. Η τεχνική αυτή αποσκοπεί στην εύκολη ανάκτηση μίας εργασίας από την τοπική μηχανή αναζήτησης χωρίς να χρειαστεί να έχεις ως όρισμα όλα τα μεταδεδομένα της. Τα μεταδεδομένα συλλέγονται ακριβώς όπως είναι ενσωματωμένα στη μη-δομημένη μορφή HTML και με λίγες παραλλαγές, ώστε να αποθηκευτεί το κύριο περιεχόμενο του κάθε πεδίου. Αξίζει να σημειωθεί ότι οι συγγραφείς και τα μαθήματα και υπο-μαθήματα που σχετίζονται με την εργασία, έχουν αποθηκευτεί σε μία δομή αντί ενός ενιαίου κειμενικού περιεχομένου για την καλύτερη οργάνωση των δεδομένων. Επίσης, κάποιες εργασίες δεν διαθέτουν περιεχόμενο στο πεδίο σχόλια και γι' αυτό αντικαθίσταται μ' ένα κενό. Τέλος, τα δεδομένα κάθε εργασίας αποθηκεύονται σε μία δομή λεξικού.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Cornell University arXiv

We gratefully acknowledge support from the Simons Foundation and member institutions.

Search... All fields Search Help | Advanced Search Login

Showing 1-100 of 221,264 results for all: Statistics

Statistics All fields Search Advanced Search

Show abstracts Hide abstracts

p.title.is_5.mathjax 820 > 21.08
Color: ■ #202020
Font: 17.5px "Open Sans", "Lucida Grande", Helvetica, sans-serif
Margin: 0px 0px 4.375px

Accessibility: Contrast: Aa 13.77
Name: paragraph
Role: Keyboard-focusable
1. **arXiv**
Clustering-based spatial interpolation of parametric post-processing models
Authors: Sándor Baran, Mária Lakatos
Abstract: Since the start of the operational use of ensemble prediction systems, ensemble-based probabilistic forecasting has become the most advanced approach in weather prediction. However, despite the persistent development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of **statistical** post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated weather quantity. Parameter estimation in these models is based on training data consisting of past forecast-observation pairs, thus post-processed forecasts are usually available only at those locations where training data are accessible. We propose a general clustering-based interpolation technique of extending calibrated predictive distributions from observation stations to any location in the ensemble domain where there are ensemble forecasts at hand. Focusing on the ensemble model output **statistics** (EMOS) post-processing technique, in a case study based on wind speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts, we demonstrate the predictive performance of various versions of the suggested method and show its superiority over the regionally estimated and interpolated EMOS models and the raw ensemble forecasts as well. ▲ [Less](#)
Submitted 25 January, 2024; originally announced January 2024.
Comments: 19 pages, 6 figures

2. arXiv:2401.14378 [pdf, ps, other] cond-mat.stat-mech

HTML View Source

DOM Elements Console Sources Network Performance

Third-party cookie phaseout
The Issues panel now warns you about the cookies that will be affected by the upcoming deprecation and phaseout of third-party cookies.

Εικόνα 1.B.1 Ανάκτηση του τίτλου της εργασίας

Cornell University arXiv

We gratefully acknowledge support from the Simons Foundation and member institutions.

Search... All fields Search Help | Advanced Search Login

Showing 1-100 of 221,264 results for all: Statistics

Statistics All fields Search Advanced Search

Show abstracts Hide abstracts

p.authors 830 > 21
Color: ■ #333333
Font: 14px "Open Sans", "Lucida Grande", Helvetica, sans-serif
Margin: 0px 0px 3.5px

Accessibility: Contrast: Aa 13.77
Name: paragraph
Role: Keyboard-focusable
1. **arXiv**
Clustering-based spatial interpolation of parametric post-processing models
Authors: Sándor Baran, Mária Lakatos
Abstract: Since the start of the operational use of ensemble prediction systems, ensemble-based probabilistic forecasting has become the most advanced approach in weather prediction. However, despite the persistent development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of **statistical** post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated weather quantity. Parameter estimation in these models is based on training data consisting of past forecast-observation pairs, thus post-processed forecasts are usually available only at those locations where training data are accessible. We propose a general clustering-based interpolation technique of extending calibrated predictive distributions from observation stations to any location in the ensemble domain where there are ensemble forecasts at hand. Focusing on the ensemble model output **statistics** (EMOS) post-processing technique, in a case study based on wind speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts, we demonstrate the predictive performance of various versions of the suggested method and show its superiority over the regionally estimated and interpolated EMOS models and the raw ensemble forecasts as well. ▲ [Less](#)
Submitted 25 January, 2024; originally announced January 2024.
Comments: 19 pages, 6 figures

2. arXiv:2401.14378 [pdf, ps, other] cond-mat.stat-mech

HTML View Source

DOM Elements Console Sources Network Performance

Third-party cookie phaseout
The Issues panel now warns you about the cookies that will be affected by the upcoming deprecation and phaseout of third-party cookies.

Εικόνα 1.B.2 Ανάκτηση των συγγραφέων της εργασίας

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arkiv

Search All fields Search

Help | Advanced Search Login

Showing 1-100 of 221,264 results for all: Statistics

Search v0.5.6 released 2020-02-24 | Feedback?

Statistics All fields Search

Show abstracts Hide abstracts

Advanced Search

div#marginless 830 x 28.67
Color ■ #333333
Font: 14px "Open Sans", "Lucida Grande", "Hei...
ACCESIBILITY
Name generic
Role
Keyboard-focusable
Announcement date (newest first) Go
1 2 3 4 5 ... Next

1. arXiv:2401.14393 [pdf, other] stat.AP stat.ME

Clustering-based spatial interpolation of parametric post-processing models

Authors: Sándor Barany, Mária Lakatos

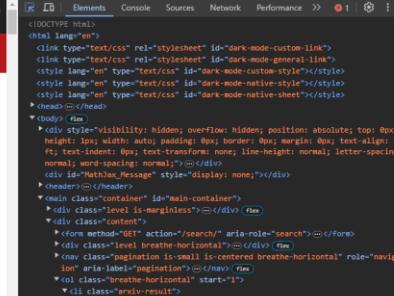
Abstract: Since the start of the operational use of ensemble prediction systems, ensemble-based probabilistic forecasting has become the most advanced approach in weather prediction. However, despite the persistent development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of **statistical** post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated weather quantity. Parameter estimation in these models is based on training data consisting of past forecast-observation pairs, thus post-processed forecasts are usually available only at those locations where training data are accessible. We propose a general clustering-based interpolation technique of extending calibrated predictive distributions from observation stations to any location in the ensemble domain where there are ensemble forecasts at hand. Focusing on the ensemble model output **statistics** (EMOS) post-processing technique, in a case study based on wind speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts, we demonstrate the predictive performance of various versions of the suggested method and show its superiority over the regionally estimated and interpolated EMOS models and the raw ensemble forecasts as well. ▾ [Less](#)

Submitted 25 January 2024, originally announced January 2024.

Comments: 19 pages, 6 figures

2. arXiv:2401.14378 [pdf, ps, other] cond-mat.stat-mech

Final file dimensions with nonlocal closure measure

HTML Elements Console Sources Network Performance > 

Εικόνα 1.B.3 Ανάκτηση των μαθημάτων που σχετίζονται με την εργασία

Εικόνα 1.B.4 Ανάκτηση της περίληψης της εργασίας

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

The screenshot shows the arXiv search interface. At the top, there's a navigation bar with links for Cornell University, arXiv, and a search bar. Below the search bar is a dropdown menu for 'All fields' and a 'Search' button. To the right of the search bar are 'Help' and 'Advanced Search' links, and a 'Login' link. The main content area displays a search result for 'Statistics'. It shows the title 'Showing 1-100 of 221,264 results for all: Statistics' and a search bar with 'v0.5.6 released 2020-02-24' and a 'Feedback?' link. Below the search bar are two buttons: 'Statistics' (selected) and 'Show abstracts' (unchecked). The search results are presented in a table format with columns for '100 results per page' (dropdown), 'Sort results by' (dropdown set to 'Announcement date (newest first)'), and a 'Go' button. The results list includes the first entry: 'arXiv:2401.14393 [pdf, other] stat.AP stat.ME Clustering-based spatial interpolation of parametric post-processing models'. This entry has a 'View details' button and a 'Comments' section. The comments table includes columns for 'p.comments.is_size-7' (checkbox checked), 'Color' (checkbox checked), 'Font' (dropdown '11.9px "Open Sans", "Lucida Grande", "Hiragino Sans GB", "Microsoft YaHei", sans-serif'), 'Margin' (dropdown '0px 0px 2.975px'), 'ACCESSIBILITY' (checkbox checked), 'Name' (dropdown 'paragraph'), 'Role' (dropdown 'Keyboard-focusable'), and 'Comments' (checkbox checked). The comments table also shows a row for 'arXiv:2401.14378 [pdf, ps, other] cond-mat.stat-mech'. The bottom of the page features a footer with links for 'arXiv:2401.14393' and 'arXiv:2401.14378'.

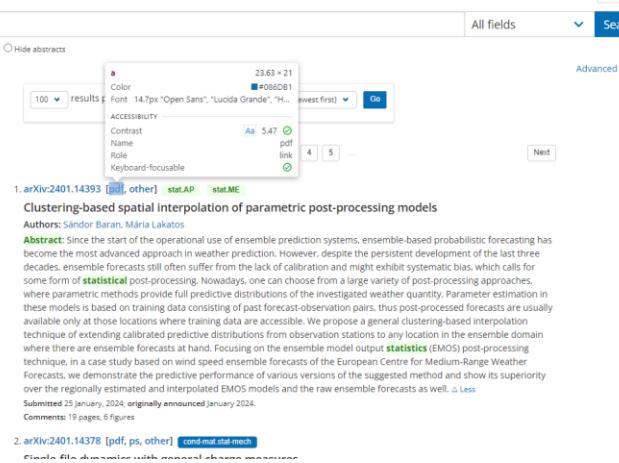
Εικόνα 1.B.5 Ανάκτηση των σχολίων της εργασίας

Εικόνα 1.B.6 Ανάκτηση της ημερομηνίας δημοσίευσης της εργασίας

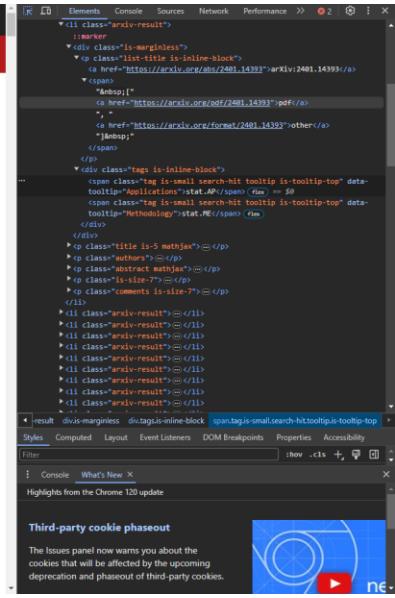
ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



Showing 1-100 of 221,264 results for all: Statistics



1. arXiv:2401.14393 [pdf, other] stat.AP stat.ME
Clustering-based spatial interpolation of parametric post-processing models
Authors: Sándor Baran, Mária Lakatos
Abstract: Since the start of the operational use of ensemble prediction systems, ensemble-based probabilistic forecasting has become the most advanced approach in weather prediction. However, despite the persistent development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of statistical post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated weather quantity. Parameter estimation in these models is based on training data consisting of past forecast-observation pairs, thus post-processed forecasts are usually available only at those locations where training data are accessible. We propose a general clustering-based interpolation technique of extending calibrated predictive distributions from observation stations to any location in the ensemble domain where there are ensemble forecasts at hand. Focusing on the ensemble model output **statistics** (EMOS) post-processing technique, in a case study based on wind speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts, we demonstrate the predictive performance of various versions of the suggested method and show its superiority over the regionally estimated and interpolated EMOS models and the raw ensemble forecasts as well. [△ Less](#)
Submitted 25 January, 2024; originally announced January 2024.
Comments: 19 pages, 6 figures

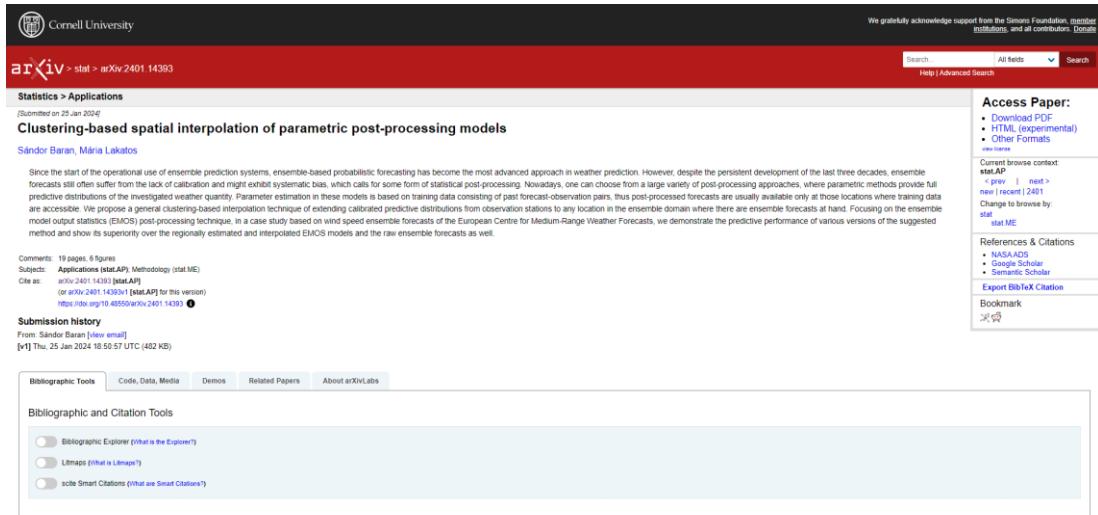


The screenshot shows the Chrome DevTools open, specifically the Elements tab, displaying the HTML structure of an arXiv search result. The code highlights several spans with the class "arxiv-result", each containing a link to a PDF or other file. The DevTools interface includes tabs for Elements, Console, Sources, Network, and Performance.

Εικόνα 1.B.7 Ανάκτηση του συνδέσμου λήψης της εργασίας σε pdf

1.β.3 Αξιολόγηση

Η αρχική υλοποίηση του σταχυολογητή ήταν πιο πολυπλόκη, καθώς, για κάθε εργασία, η συλλογή των μεταδεδομένων (web scrape) λάμβανε χώρα στις ξεχωριστές σελίδες της κάθε εργασίας με σύνδεσμο arxiv.org/abs/XXXX.XXXXX (π.χ. <https://arxiv.org/abs/2401.14393>). Αυτό δημιουργούσε σημαντικές καθυστερήσεις στην απόδοση του σταχυολογητή, καθώς, η συλλογή δεδομένων από 200-800 εργασίες σήμανε την αποστολή HTTP-GET αιτήματος σε 200-800 συνδέσμους εργασιών για την εμφάνιση των αντίστοιχων σελίδων με τα δεδομένα της κάθε εργασίας σε μη-δομημένη μορφή HTML. Η τελική υλοποίηση μειώνει την ανάκτηση σελιδών από 200-800 συνδέσμους σε 2-8 συνδέσμους, όπου τα δεδομένα των εργασιών συλλέγονται αυτούσια όπως είναι ενσωματωμένα στη μη-δομημένη μορφή HTML και αποθηκεύονται καθαρογραμμένα σύμφωνα με τις απαραίτητες τεχνικές, σε μία δομή δεδομένων, επιτυγχάνοντας με αυτό τον τρόπο στο αξιόπιστο μέγεθος του αποθετηρίου (dataset) και στην απόδοση του σταχυολογητή.



arXiv > stat > arXiv:2401.14393
Statistics > Applications
[Submitted on 25 Jan 2024]
Clustering-based spatial interpolation of parametric post-processing models
Sándor Baran, Mária Lakatos
Abstract: Since the start of the operational use of ensemble prediction systems, ensemble-based probabilistic forecasting has become the most advanced approach in weather prediction. However, despite the persistent development of the last three decades, ensemble forecasts still often suffer from the lack of calibration and might exhibit systematic bias, which calls for some form of statistical post-processing. Nowadays, one can choose from a large variety of post-processing approaches, where parametric methods provide full predictive distributions of the investigated weather quantity. Parameter estimation in these models is based on training data consisting of past forecast-observation pairs, thus post-processed forecasts are usually available only at those locations where training data are accessible. We propose a general clustering-based interpolation technique of extending calibrated predictive distributions from observation stations to any location in the ensemble domain where there are ensemble forecasts at hand. Focusing on the ensemble model output **statistics** (EMOS) post-processing technique, in a case study based on wind speed ensemble forecasts of the European Centre for Medium-Range Weather Forecasts, we demonstrate the predictive performance of various versions of the suggested method and show its superiority over the regionally estimated and interpolated EMOS models and the raw ensemble forecasts as well.
Comments: 19 pages, 6 figures
Subjects: Applications (stat.AP); Methodology (stat.ME)
Cite as: arXiv:2401.14393 [stat.AP]
(or arXiv:2401.14393v1 [stat.AP] for this version)
<https://doi.org/10.48550/arXiv.2401.14393>
Submission history
From: Sándor Baran [view email]
[v1] Tue, 25 Jan 2024 18:50:57 UTC (482 KB)
Bibliographic Tools
Code, Data, Media Demos Related Papers About arXiv.labs
Bibliographic and Citation Tools
 Bibliographic Explorer ([What is the Explorer?](#))
 Lmaps ([What is Lmaps?](#))
 smart Citations ([What are Smart Citations?](#))

Εικόνα 1.B.8 Η σελίδα της εργασίας με όλα τα δεδομένα

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

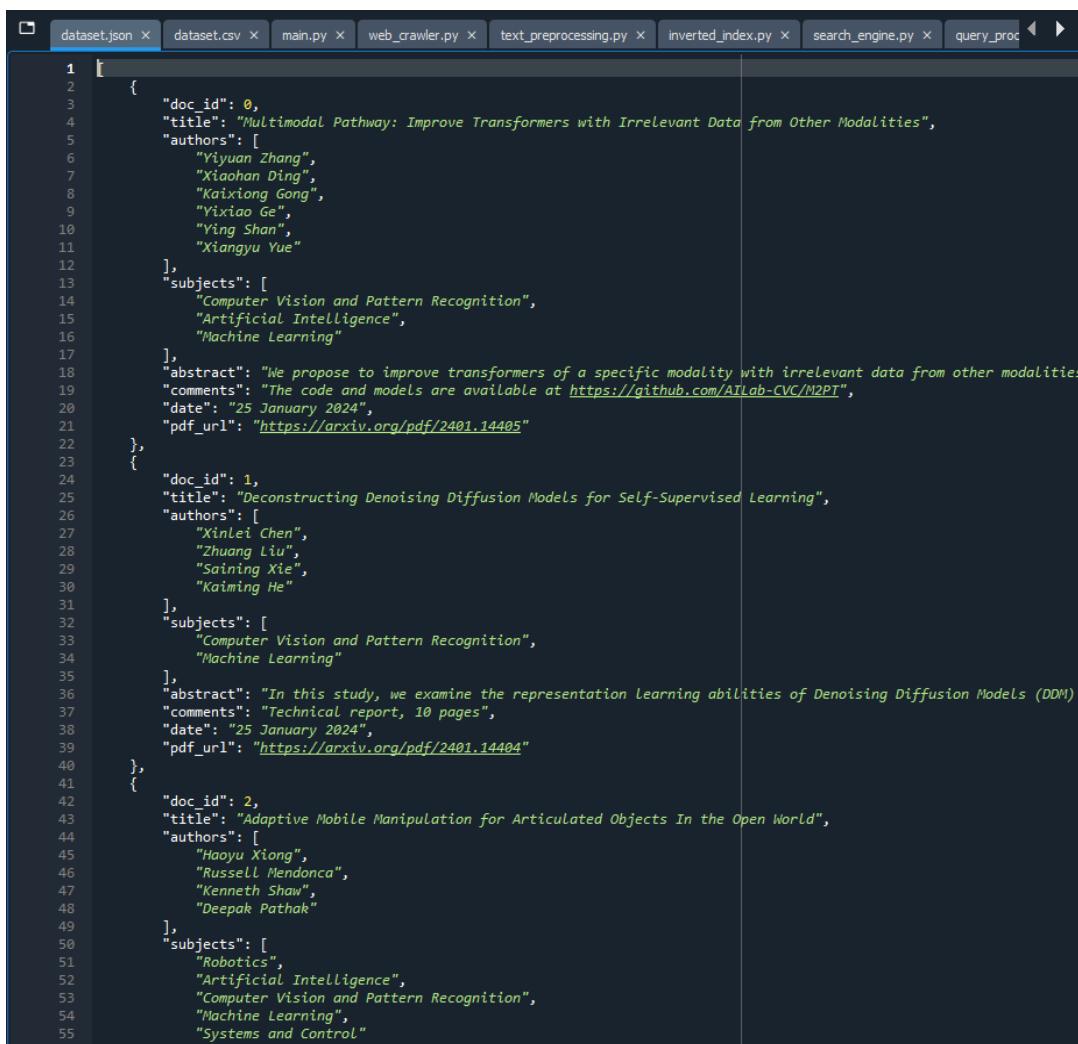
1.γ. Αποθήκευση δεδομένων σε δομημένη μορφή

1.γ.1 Σχεδιασμός

Ο σχεδιασμός υιοθετεί την δομημένη μορφή JSON αντί της μορφής CSV για την αποθήκευση των δεδομένων των εργασιών και με την εκτέλεση αυτού του βήματος, ολοκληρώνεται η δημιουργία του αποθετηρίου (dataset) της τοπικής μηχανής αναζήτησης

1.γ.2 Υλοποίηση

Η συλλογή των μεταδεδομένων των εργασίων από την μη-δομημένη μορφή HTML έχει ολοκληρωθεί στην [υλοποίηση](#) του βήματος 1.β. Υλοποίηση web crawler, οπότε αυτό που απομένει είναι να εγγραφούν τα δεδομένα σ' ένα .json αρχείο και το αποθετήριο (dataset) της τοπική μηχανής αναζήτησης να είναι αυτό. Η υλοποίηση λαμβάνει χώρα στο κύριο πρόγραμμα (main.py), όπου τα αποτελέσματα που επέστρεψε ο σταχυολογητής σε μορφή λεξικού αποθηκεύονται σε JSON μορφή.



The screenshot shows a code editor with multiple tabs at the top: dataset.json, dataset.csv, main.py, web_crawler.py, text_preprocessing.py, inverted_index.py, search_engine.py, and query_proc. The dataset.json tab is active, displaying a JSON array of three documents. Each document has fields: doc_id, title, authors, subjects, abstract, comments, date, and pdf_url. The first document is about 'Multimodal Pathway: Improve Transformers with Irrelevant Data from Other Modalities'. The second is about 'Deconstructing Denoising Diffusion Models for Self-Supervised Learning'. The third is about 'Adaptive Mobile Manipulation for Articulated Objects In the Open World'. The JSON code is numbered from 1 to 55.

```
[{"doc_id": 0, "title": "Multimodal Pathway: Improve Transformers with Irrelevant Data from Other Modalities", "authors": ["Yiyuan Zhang", "Xiaohan Ding", "Kaixiong Gong", "Vixiao Ge", "Ying Shan", "Xiangyu Yue"], "subjects": ["Computer Vision and Pattern Recognition", "Artificial Intelligence", "Machine Learning"], "abstract": "We propose to improve transformers of a specific modality with irrelevant data from other modalities", "comments": "The code and models are available at https://github.com/AIILab-CVC/M2PT", "date": "25 January 2024", "pdf_url": "https://arxiv.org/pdf/2401.14405"}, {"doc_id": 1, "title": "Deconstructing Denoising Diffusion Models for Self-Supervised Learning", "authors": ["Xinlei Chen", "Zhuang Liu", "Saining Xie", "Kaiming He"], "subjects": ["Computer Vision and Pattern Recognition", "Machine Learning"], "abstract": "In this study, we examine the representation learning abilities of Denoising Diffusion Models (DDM)", "comments": "Technical report, 10 pages", "date": "25 January 2024", "pdf_url": "https://arxiv.org/pdf/2401.14404"}, {"doc_id": 2, "title": "Adaptive Mobile Manipulation for Articulated Objects In the Open World", "authors": ["Haoyu Xiong", "Russell Hendonca", "Kenneth Shaw", "Deepak Pathak"], "subjects": ["Robotics", "Artificial Intelligence", "Computer Vision and Pattern Recognition", "Machine Learning", "Systems and Control"]}],
```

Εικόνα 1.Γ.1 Το αποθετήριο σε μορφή JSON

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

1.γ.3 Αξιολόγηση

Η επιλογή της μορφής JSON αντί της CSV για την αποθήκευση των δεδομένων των εργασιών σε δομημένη μορφή, αιτιολογείται από το γεγονός ότι υπάρχουν δεδομένα (συγγραφείς, μαθήματα) τα οποία είναι πιο ευανάγνωστα σε μία δομή (π.χ. λίστα) που υποστηρίζει το JSON παρά σ' ένα ενιαίο κείμενο το οποίο διαχωρίζεται με κόμμα (CSV). Ο όγκος πληροφορίας κυμαίνεται από 200 εώς 800 εργασίες, όπου κάθε εργασία περιλαμβάνει 8 πεδία από δεδομένα (τεκμηρίωση στην [υλοποίηση](#) του βήματος 1.β. Υλοποίηση web-crawler). Συνεπώς, η αξιοποίηση του μέγαλου όγκου δεδομένων είναι πιο εύκολα διαχειρίσιμη σε μορφή JSON παρά σε απλή μορφή CSV.

Εικόνα 1.Γ.2 Το αποθετήριο σε μορφή CSV

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

2. Προεπεξεργασία κειμένου (Text Processing)

2.α. Σχεδιασμός

Ο σχεδιασμός ακολουθεί την προεπεξεργασία του κειμενικού περιεχομένου (πεδίο Abstract) του αποθετηρίου που δημιουργήθηκε στο [βήμα 1](#), με διάφορες τεχνικές προεπεξεργασίας κειμένου. Η προεπεξεργασία κειμένου είναι μία από τις πιο βασικές τεχνικές επεξεργασίας της φυσικής γλώσσας (NLP), όπου επιτυγχάνει την αντιστοίχιση όρων των δεδομένων με το ερώτημα αναζήτησης, ώστε η ασάφεια που προφανώς υφίσταται να υπάρχει στην διατύπωση του ερωτήματος χρήστη, να μην επηρεάζει το αποτέλεσμα αναζήτησης. Η προεπεξεργασία σχεδιάστηκε με απότερο σκοπό να ομαδοποιήσει διάφορες εκδοχές διατύπωσης ενός ερωτήματος χρήστη, ώστε να μην υπάρχει μεγάλη απόκλιση στα αποτελέσματα αναζήτησης, όταν πρόκειται για το ίδιο επιθυμητό αποτέλεσμα αλλά να επιδιώκεται με διαφορετική διατύπωση. Ο σχεδιασμός ακολουθεί διάφορες τεχνικές προεπεξεργασίας κειμένου για την προεπεξεργασία των δεδομένων των εργασιών του αποθετηρίου. Οι τεχνικές που σχεδιάστηκαν για υλοποίηση είναι οι εξής:

- Tokenization: Ο διαχωρισμός του κειμένου σε λεκτικές μονάδες
- Punctuation characters removal: Αφαίρεση όλων των σημείων στίξης
- Special characters removal: Αφαίρεση όλων των ειδικών συμβόλων
- Normalization: Η αντικατάσταση των κεφαλαίων γραμμάτων με τα αντίστοιχα πεζά τους
- Stopwords removal: Αφαίρεση των «απαγορευμένων» λέξεων
- Stemming: Αφαίρεση καταλήξεων και διατήρηση του κύριου στέλεχους του όρου

2.β. Υλοποίηση

Οι υλοποιήσεις των τεχνικών προεπεξεργασίας κειμένου λαμβάνουν χώρα σε ξεχωριστό module (text_preprocessing.py) και είναι οι εξής:

- Tokenization: Χρήση της nltk.word_tokenize() για τον διαχωρισμό του κειμένου σε λεκτικές μονάδες και διαχωρισμός των σημείων στίξης
- Punctuation characters removal: Αφαίρεση όλων των όρων που ανήκουν στο σύνολο string.punctuation, δηλαδή, των σημείων στίξης.
- Special characters removal: Αφαίρεση όλων των όρων που δεν αποτελούν νούμερα ή γράμματα
- Normalization: Αντικατάσταση των κεφαλαίων γραμμάτων με τα αντίστοιχα πεζά τους με χρήση της ρουτίνας .lower()
- Stopwords removal: Αφαίρεση όλων των όρων που ανήκουν στο σύνολο stopwords.words('english') από το nltk.corpus. Συνήθως, πρόκειται για όρους που εμφανίζονται συχνά σε κείμενα όπως είναι τα άρθρα, οι αντωνυμίες, τα επιρρήματα κλπ.
- Stemming: Αφαίρεση καταλήξεων και διατήρηση του κύριου στέλεχους του όρου με χρήση του αλγορίθμου PorterStemmer() της nltk.stem

Δεδομένου, ότι για την αναζήτηση εργασίων θα χρειαστεί και ένα ερώτημα αναζήτησης (query), θα πρέπει και αυτό να προεπεξεργαστεί (λεπτομέρειες στο [βήμα 4 Μηχανή αναζήτησης](#)). Στα ερωτήματα αναζήτησης που

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

περιλαμβάνουν Boolean πράξεις δεν πραγματοποιείται η τεχνική προεπεξεργασίας «Stopwords removal», καθώς έτσι θα αφαιρεθούν και οι λογικοί τελεστές AND, NOT, OR από το ερώτημα (λεπτομέριες στο [βήμα 4.β.1 Boolean Retrieval](#)).

2.γ. Αξιολόγηση

Η προεπεξεργασία γίνεται με την σειρά που παρουσιάστηκε στον [σχεδιασμό](#) και την [υλοποίηση](#). Η αξιολόγηση βασίζεται στην επιλογή των τεχνικών προεπεξεργασίας κειμένου, καθώς, και στην σειρά οι οποίες λαμβάνουν χώρα. Οι αξιολογήσεις των λειτουργιών της κάθε τεχνικής είναι οι εξής:

- Tokenization: Η διαχώριση του κειμένου σε λεκτικές μονάδες συμβάλλει στην διαχείριση των λέξεων και των σημείων στίξης και γι' αυτό εφαρμόζεται 1^η.
- Punctuation characters removal: Η αφαίρεση των σημείων στίξης βοηθάει στον καθαρισμό του κειμένου από περιττές πληροφορίες και εφαρμόζεται 2^η γι' αυτό τον λόγο. Ωστόσο, η τεχνική αυτή ενδέχεται να επηρεάσει την κατανόηση ενός κειμένου που χρησιμοποιεί σημεία στίξης για την διάκριση όρων (π.χ. με αφαίρεση των σημείων στίξης σε μία ημερομηνία 19/03/2001, η προεπεξεργασμένη ημερομηνία απεικονίζεται ως 3 νούμερα 19 03 2001 με αποτέλεσμα να υπάρχει μία απώλεια πληροφορίας)
- Special characters removal: Η αφαίρεση ειδικών συμβόλων βοηθάει επίσης στον καθαρισμό του κειμένου από περιττές πληροφορίες. Ωστόσο, όπως το ίδιο ισχύει και με την τεχνική punctuation characters removal, ενδέχεται να επηρεάσει την κατανόηση του κειμένου (π.χ. με αφαίρεση των ειδικών συμβόλων σ' ένα email jondoe@gmail.com, το προεπεξεργασμένο email απεικονίζεται ως jondoe gmail com, με αποτέλεσμα να υπάρχει απώλεια πληροφορίας)
- Normalization: Η κανονικοποίηση βοηθάει στο να φέρει τους όρους σε μία κοινή μορφή, μετατρέποντας όλα τα κεφαλαία γράμματα σε πεζά.
- Stopwords removal: Η αφαίρεση των απαγορευμένων λέξεων καθαρίζει το κείμενο από λέξεις που εμφανίζονται πολύ συχνά και δεν προσφέρουν σημαντική πληροφορία. Ωστόσο, ενδέχεται να επηρεάσει την κατανόηση του κειμένου σε όρους που συνοδεύονται με απαγορευμένες λέξεις (π.χ. με αφαίρεση των απαγορευμένων λέξεων στο King of Denmark, το προεπεξεργασμένο κείμενο απεικονίζεται ως king denmark, με αποτέλεσμα να χάνεται πληροφορία).
- Stemming: Η διαδικασία «stemming» αφαιρεί τις καταλήξεις από τις λέξεις και κρατάει το κύριο στέλεχος της λέξης επιτυγχάνοντας με αυτό τον τρόπο να ομαδοποιήσει διάφορες εκδοχές μίας λέξης σε ένα κοινό στέλεχος. Ωστόσο και εδώ υπάρχει ο κίνδυνος απώλειας πληροφορίας (π.χ. στα boolean ερωτήματα αναζήτησης operational AND research, operating AND system, operative AND dentistry, η διαδικασία stemming ομαδοποιεί τους όρους operational, operating, operative στο κοινό στέλεχος oper, με αποτέλεσμα να υπάρχει απώλεια πληροφορίας)

Η τεχνική lemmatization προσφέρει πιο ακριβή αποτελέσματα, καθώς λαμβάνει υπόψιν τη γραμματική της λέξης για την μετατροπή της στη βασική της μορφή σε αντίθεση με την τεχνική stemming που αφαιρεί απλά καταλήξεις από λέξεις και παρουσιάζει ενδεχόμενο πρόβλημα απώλειας πληροφορίας. Ωστόσο, με δεδομένο ότι το αποθετήριο (dataset) είναι μικρό σε σύγκριση με αποθετήρια μεγάλων μηχανών αναζήτησης (αποτελούμενο από εκατομμύρια έγγραφα), η τεχνική stemming προσφέρει απλότητα και ταχύτητα στην απόδοση της μηχανής αναζήτησης και αυτός είναι ο λόγος που προτιμήθηκε ενάντι της τεχνικής lemmatization.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

2.δ. Εφαρμογή

Για τις ανάγκες της εφαρμογής της προεπεξεργασίας κειμένου χρησιμοποιήθηκε το πεδίο «abstract» της εργασίας με τα παρακάτω δεδομένα:

{

 "doc_id": 0,

 "title": "Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration",

 "authors": [

 "M. Reichert",

 "M. Bugli",

 "J. Guilet",

 "M. Obergaulinger",

 "M. \u00c1lvaro Aloy",

 "A. Arcones"

],

 "subjects": [

 "High Energy Astrophysical Phenomena"

],

 "abstract": "The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, \$\u03b2^{\{-\}}\$-decays and \$\u03b2^{\{-\}}\$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.",

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
"comments": " ",  
"date": "25 January 2024",  
"pdf_url": https://arxiv.org/pdf/2401.14402  
}
```

```
In [41]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')  
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing  
----- Tokenization -----  
['The', 'production', 'of', 'heavy', 'elements', 'is', 'one', 'of', 'the', 'main', 'by-products', 'of', 'the', 'explosive', 'end',  
'of', 'massive', 'stars', '.', 'A', 'long', 'sought', 'goal', 'is', 'finding', 'differentiated', 'patterns', 'in', 'the',  
'nucleosynthesis', 'yields', ',', 'which', 'could', 'permit', 'identifying', 'a', 'number', 'of', 'properties', 'of', 'the',  
'explosive', 'core', '.', 'Among', 'them', ',', 'the', 'traces', 'of', 'the', 'magnetic', 'field', 'topology', 'are',  
'particularly', 'important', 'for', '\\emph', '{}', 'extreme', 'supernova', 'explosions', ',', 'most', 'likely', 'hosted',  
'by', 'magnetorotational', 'effects', '.', 'We', 'investigate', 'the', 'nucleosynthesis', 'of', 'five', 'state-of-the-art',  
'magnetohydrodynamic', 'models', 'with', 'fast', 'rotation', 'that', 'have', 'been', 'previously', 'calculated', 'in', 'full',  
'3D', 'and', 'that', 'involve', 'an', 'accurate', 'neutrino', 'transport', '(', 'M1', ')', 'One', 'of', 'the', 'models',  
'does', 'not', 'contain', 'any', 'magnetic', 'field', 'and', 'synthesizes', 'elements', 'around', 'the', 'iron', 'group', ',',  
'in', 'agreement', 'with', 'other', 'CC-SNe', 'models', 'in', 'literature', 'All', 'other', 'models', 'host', 'a', 'strong',  
'magnetic', 'field', 'of', 'the', 'same', 'intensity', ',', 'but', 'with', 'different', 'topology', 'For', 'the', 'first',  
'time', 'we', 'investigate', 'the', 'nucleosynthesis', 'of', 'MR-SNe', 'models', 'with', 'a', 'quadrupolar', 'magnetic',  
'field', 'and', 'a', '90', 'degree', 'tilted', 'dipole', 'We', 'obtain', 'a', 'large', 'variety', 'of', 'ejecta',  
'compositions', 'reaching', 'from', 'iron', 'nuclei', 'to', 'nuclei', 'up', 'to', 'the', 'third', 'r-process', 'peak', '.', 'We',  
'assess', 'the', 'robustness', 'of', 'our', 'results', 'by', 'considering', 'the', 'impact', 'of', 'different', 'nuclear',  
'physics', 'uncertainties', 'such', 'as', 'different', 'nuclear', 'masses', '$', 'B^', '{}', '$', '-decays', 'and',  
'$', 'B^', '{}', '$', '-delayed', 'neutron', 'emission', 'probabilities', '$', 'neutrino', 'reactions', '$', 'fission',  
'$', 'and', 'a', 'feedback', 'of', 'nuclear', 'energy', 'on', 'the', 'temperature', 'We', 'find', 'that', 'the',  
'qualitative', 'results', 'do', 'not', 'change', 'with', 'different', 'nuclear', 'physics', 'input', 'The', 'properties',  
'of', 'the', 'explosion', 'dynamics', 'and', 'the', 'magnetic', 'field', 'configuration', 'are', 'the', 'dominant', 'factors',  
'determining', 'the', 'ejecta', 'composition', '.']
```

Εικόνα 2.Δ.1 Tokenization

```
In [43]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')  
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing  
----- Punctuation characters removal -----  
['The', 'production', 'of', 'heavy', 'elements', 'is', 'one', 'of', 'the', 'main', 'by-products', 'of', 'the', 'explosive', 'end',  
'of', 'massive', 'stars', 'A', 'long', 'sought', 'goal', 'is', 'finding', 'differentiated', 'patterns', 'in', 'the',  
'nucleosynthesis', 'yields', 'which', 'could', 'permit', 'identifying', 'a', 'number', 'of', 'properties', 'of', 'the',  
'explosive', 'core', 'Among', 'them', 'the', 'traces', 'of', 'the', 'magnetic', 'field', 'topology', 'are', 'particularly',  
'important', 'for', '\\emph', '{}', 'extreme', 'supernova', 'explosions', 'most', 'likely', 'hosted', 'by', 'magnetorotational',  
'effects', 'We', 'investigate', 'the', 'nucleosynthesis', 'of', 'five', 'state-of-the-art', 'magnetohydrodynamic', 'models',  
'with', 'fast', 'rotation', 'that', 'have', 'been', 'previously', 'calculated', 'in', 'full', '3D', 'and', 'that', 'involve', 'an',  
'accurate', 'neutrino', 'transport', 'M1', 'One', 'of', 'the', 'models', 'does', 'not', 'contain', 'any', 'magnetic', 'field',  
'and', 'synthesizes', 'elements', 'around', 'the', 'iron', 'group', 'in', 'agreement', 'with', 'other', 'CC-SNe', 'models', 'in',  
'literature', 'All', 'other', 'models', 'host', 'a', 'strong', 'magnetic', 'field', 'of', 'the', 'same', 'intensity', 'but',  
'with', 'different', 'topology', 'For', 'the', 'first', 'time', 'we', 'investigate', 'the', 'nucleosynthesis', 'of', 'MR-SNe',  
'models', 'with', 'a', 'quadrupolar', 'magnetic', 'field', 'and', 'a', '90', 'degree', 'tilted', 'dipole', 'We', 'obtain', 'a',  
'large', 'variety', 'of', 'ejecta', 'compositions', 'reaching', 'from', 'iron', 'nuclei', 'to', 'nuclei', 'up', 'to', 'the',  
'third', 'r-process', 'peak', 'We', 'assess', 'the', 'robustness', 'of', 'our', 'results', 'by', 'considering', 'the', 'impact',  
'of', 'different', 'nuclear', 'physics', 'uncertainties', 'such', 'as', 'different', 'nuclear', 'masses', 'B^', '-decays', 'and',  
'$', 'B^', '-delayed', 'neutron', 'emission', 'probabilities', '$', 'neutrino', 'reactions', 'fission', 'and', 'a', 'feedback', 'of',  
'nuclear', 'energy', 'on', 'the', 'temperature', 'We', 'find', 'that', 'the', 'qualitative', 'results', 'do', 'not', 'change',  
'with', 'different', 'nuclear', 'physics', 'input', 'The', 'properties', 'of', 'the', 'explosion', 'dynamics', 'and', 'the',  
'magnetic', 'field', 'configuration', 'are', 'the', 'dominant', 'factors', 'determining', 'the', 'ejecta', 'composition']
```

Εικόνα 2.Δ.2 Punctuation characters removal

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
In [45]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing
----- Special characters removal -----
['The', 'production', 'of', 'heavy', 'elements', 'is', 'one', 'of', 'the', 'main', 'byproducts', 'of', 'the', 'explosive', 'end',
'of', 'massive', 'stars', 'A', 'long', 'sought', 'goal', 'is', 'finding', 'differentiated', 'patterns', 'in', 'the',
'nucleosynthesis', 'yields', 'which', 'could', 'permit', 'identifying', 'a', 'number', 'of', 'properties', 'of', 'the',
'explosive', 'core', 'Among', 'them', 'the', 'traces', 'of', 'the', 'magnetic', 'field', 'topology', 'are', 'particularly',
'important', 'for', 'emph', 'extreme', 'supernova', 'explosions', 'most', 'likely', 'hosted', 'by', 'magnetorotational', 'effects',
'We', 'investigate', 'the', 'nucleosynthesis', 'of', 'five', 'stateoftheheart', 'magnetohydrodynamic', 'models', 'with', 'fast',
'rotation', 'that', 'have', 'been', 'previously', 'calculated', 'in', 'full', '3D', 'and', 'that', 'involve', 'an', 'accurate',
'neutrino', 'transport', 'M1', 'One', 'of', 'the', 'models', 'does', 'not', 'contain', 'any', 'magnetic', 'field', 'and',
'synthesizes', 'elements', 'around', 'the', 'iron', 'group', 'in', 'agreement', 'with', 'other', 'CCSNe', 'models', 'in',
'literature', 'All', 'other', 'models', 'host', 'a', 'strong', 'magnetic', 'field', 'of', 'the', 'same', 'intensity', 'but',
'with', 'different', 'topology', 'For', 'the', 'first', 'time', 'we', 'investigate', 'the', 'nucleosynthesis', 'of', 'MRSNe',
'models', 'with', 'a', 'quadrupolar', 'magnetic', 'field', 'and', 'a', '90', 'degree', 'tilted', 'dipole', 'We', 'obtain', 'a',
'large', 'variety', 'of', 'ejecta', 'compositions', 'reaching', 'from', 'iron', 'nuclei', 'to', 'nuclei', 'up', 'to', 'the',
'third', 'rprocess', 'peak', 'We', 'assess', 'the', 'robustness', 'of', 'our', 'results', 'by', 'considering', 'the', 'impact',
'of', 'different', 'nuclear', 'physics', 'uncertainties', 'such', 'as', 'different', 'nuclear', 'masses', 'decays', 'and',
'delayed', 'neutron', 'emission', 'probabilities', 'neutrino', 'reactions', 'fission', 'and', 'a', 'feedback', 'of', 'nuclear',
'energy', 'on', 'the', 'temperature', 'We', 'find', 'that', 'the', 'qualitative', 'results', 'do', 'not', 'change', 'with',
'different', 'nuclear', 'physics', 'input', 'The', 'properties', 'of', 'the', 'explosion', 'dynamics', 'and', 'the', 'magnetic',
'field', 'configuration', 'are', 'the', 'dominant', 'factors', 'determining', 'the', 'ejecta', 'composition']
```

Εικόνα 2.Δ.3 Special characters removal

```
In [47]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing
----- Normalization -----
['the', 'production', 'of', 'heavy', 'elements', 'is', 'one', 'of', 'the', 'main', 'byproducts', 'of', 'the', 'explosive', 'end',
'of', 'massive', 'stars', 'a', 'long', 'sought', 'goal', 'is', 'finding', 'differentiated', 'patterns', 'in', 'the',
'nucleosynthesis', 'yields', 'which', 'could', 'permit', 'identifying', 'a', 'number', 'of', 'properties', 'of', 'the',
'explosive', 'core', 'among', 'them', 'the', 'traces', 'of', 'the', 'magnetic', 'field', 'topology', 'are', 'particularly',
'important', 'for', 'emph', 'extreme', 'supernova', 'explosions', 'most', 'likely', 'hosted', 'by', 'magnetorotational', 'effects',
've', 'investigate', 'the', 'nucleosynthesis', 'of', 'five', 'stateoftheheart', 'magnetohydrodynamic', 'models', 'with', 'fast',
'rotation', 'that', 'have', 'been', 'previously', 'calculated', 'in', 'full', '3d', 'and', 'that', 'involve', 'an', 'accurate',
'neutrino', 'transport', 'm1', 'one', 'of', 'the', 'models', 'does', 'not', 'contain', 'any', 'magnetic', 'field', 'and',
'synthesizes', 'elements', 'around', 'the', 'iron', 'group', 'in', 'agreement', 'with', 'other', 'ccsne', 'models', 'in',
'literature', 'all', 'other', 'models', 'host', 'a', 'strong', 'magnetic', 'field', 'of', 'the', 'same', 'intensity', 'but',
'with', 'different', 'topology', 'for', 'the', 'first', 'time', 'we', 'investigate', 'the', 'nucleosynthesis', 'of', 'mrsne',
'models', 'with', 'a', 'quadrupolar', 'magnetic', 'field', 'and', 'a', '90', 'degree', 'tilted', 'dipole', 'we', 'obtain', 'a',
'large', 'variety', 'of', 'ejecta', 'compositions', 'reaching', 'from', 'iron', 'nuclei', 'up', 'to', 'the',
'third', 'rprocess', 'peak', 'we', 'assess', 'the', 'robustness', 'of', 'our', 'results', 'by', 'considering', 'the', 'impact',
'of', 'different', 'nuclear', 'physics', 'uncertainties', 'such', 'as', 'different', 'nuclear', 'masses', 'decays', 'and',
'delayed', 'neutron', 'emission', 'probabilities', 'neutrino', 'reactions', 'fission', 'and', 'a', 'feedback', 'of', 'nuclear',
'energy', 'on', 'the', 'temperature', 'we', 'find', 'that', 'the', 'qualitative', 'results', 'do', 'not', 'change', 'with',
'different', 'nuclear', 'physics', 'input', 'the', 'properties', 'of', 'the', 'explosion', 'dynamics', 'and', 'the', 'magnetic',
'field', 'configuration', 'are', 'the', 'dominant', 'factors', 'determining', 'the', 'ejecta', 'composition']
```

Εικόνα 2.Δ.4 Normalization

```
In [49]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing
----- Stop-words removal -----
['production', 'heavy', 'elements', 'one', 'main', 'byproducts', 'explosive', 'end', 'massive', 'stars', 'long', 'sought', 'goal',
'finding', 'differentiated', 'patterns', 'nucleosynthesis', 'yields', 'could', 'permit', 'identifying', 'number', 'properties',
'explosive', 'core', 'among', 'traces', 'magnetic', 'field', 'topology', 'particularly', 'important', 'emph', 'extreme',
'supernova', 'explosions', 'likely', 'hosted', 'magnetorotational', 'effects', 'investigate', 'nucleosynthesis', 'five',
'stateoftheheart', 'magnetohydrodynamic', 'models', 'fast', 'rotation', 'previously', 'calculated', 'full', '3d', 'involve',
'accurate', 'neutrino', 'transport', 'm1', 'one', 'models', 'contain', 'magnetic', 'field', 'synthesizes', 'elements', 'around',
'iron', 'group', 'agreement', 'ccsne', 'models', 'literature', 'models', 'host', 'strong', 'magnetic', 'field', 'intensity',
'different', 'topology', 'first', 'time', 'investigate', 'nucleosynthesis', 'mrsne', 'models', 'quadrupolar', 'magnetic', 'field',
'90', 'degree', 'tilted', 'dipole', 'obtain', 'large', 'variety', 'ejecta', 'compositions', 'reaching', 'iron', 'nuclei', 'nuclei',
'third', 'rprocess', 'peak', 'assess', 'robustness', 'results', 'considering', 'impact', 'different', 'nuclear', 'physics',
'uncertainties', 'different', 'nuclear', 'masses', 'decays', 'delayed', 'neutron', 'emission', 'probabilities', 'neutrino',
'reactions', 'fission', 'feedback', 'nuclear', 'energy', 'temperature', 'find', 'qualitative', 'results', 'change', 'different',
'nuclear', 'physics', 'input', 'properties', 'explosion', 'dynamics', 'magnetic', 'field', 'configuration', 'dominant', 'factors',
'determining', 'ejecta', 'composition']
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Εικόνα 2.Δ.5 Stop-words removal

```
In [51]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing
----- Stemming -----
['product', 'heavi', 'element', 'one', 'main', 'byproduct', 'explos', 'end', 'massiv', 'star', 'long', 'sought', 'goal', 'find', 'differenti', 'pattern', 'nucleosynthesi', 'yield', 'could', 'permit', 'identifi', 'number', 'properiti', 'explos', 'core', 'among', 'trace', 'magnet', 'field', 'topolog', 'particularli', 'import', 'emph', 'extrem', 'supernova', 'explos', 'like', 'host', 'magnetotor', 'effect', 'investig', 'nucleosynthesi', 'five', 'stateofheart', 'magnetohydrodynam', 'model', 'fast', 'rotat', 'previous', 'calcul', 'full', '3d', 'involv', 'accur', 'neutrino', 'transport', 'ml', 'one', 'model', 'contain', 'magnet', 'field', 'synthes', 'element', 'around', 'iron', 'group', 'agreement', 'ccsne', 'model', 'literatur', 'model', 'host', 'strong', 'magnet', 'field', 'intens', 'differ', 'topolog', 'first', 'time', 'investig', 'nucleosynthesi', 'mrsne', 'model', 'quadrupolar', 'magnet', 'field', '90', 'degre', 'tilt', 'dipol', 'obtain', 'larg', 'variety', 'ejecta', 'composit', 'reach', 'iron', 'nuclei', 'nuclei', 'third', 'rprocess', 'peak', 'assess', 'robust', 'result', 'consid', 'impact', 'differ', 'nuclear', 'physic', 'uncertainti', 'differ', 'nuclear', 'mass', 'decay', 'delay', 'neutron', 'emiss', 'probabl', 'neutrino', 'reaction', 'fission', 'feedback', 'nuclear', 'energi', 'temperatur', 'find', 'qualit', 'result', 'chang', 'differ', 'nuclear', 'physic', 'input', 'properti', 'explos', 'dynam', 'magnet', 'field', 'configur', 'domin', 'factor', 'determin', 'ejecta', 'composit']
```

Εικόνα 2.Δ.6 Stemming (αντί Lemmatization)

```
In [53]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, inverted_index, query_processing, ranking, search_engine, text_preprocessing
----- Lemmatization -----
['production', 'heavy', 'element', 'one', 'main', 'byproduct', 'explosive', 'end', 'massive', 'star', 'long', 'sought', 'goal', 'finding', 'differentiated', 'pattern', 'nucleosynthesis', 'yield', 'could', 'permit', 'identifying', 'number', 'property', 'explosive', 'core', 'among', 'trace', 'magnetic', 'field', 'topology', 'particularly', 'important', 'emph', 'extreme', 'supernova', 'explosion', 'likely', 'hosted', 'magnetorotational', 'effect', 'investigate', 'nucleosynthesis', 'five', 'stateofheart', 'magnetohydrodynamic', 'model', 'fast', 'rotation', 'previously', 'calculated', 'full', '3d', 'involve', 'accurate', 'neutrino', 'transport', 'ml', 'one', 'model', 'contain', 'magnetic', 'field', 'synthesizes', 'element', 'around', 'iron', 'group', 'agreement', 'ccsne', 'model', 'literature', 'model', 'host', 'strong', 'magnetic', 'field', 'intensity', 'different', 'topology', 'first', 'time', 'investigate', 'nucleosynthesis', 'mrsne', 'model', 'quadrupolar', 'magnetic', 'field', '90', 'degree', 'tilted', 'dipole', 'obtain', 'large', 'variety', 'ejecta', 'composition', 'reaching', 'iron', 'nucleus', 'nucleus', 'third', 'rprocess', 'peak', 'ass', 'robustness', 'result', 'considering', 'impact', 'different', 'nuclear', 'physic', 'uncertainty', 'different', 'nuclear', 'mass', 'decay', 'delayed', 'neutron', 'emission', 'probability', 'neutrino', 'reaction', 'fission', 'feedback', 'nuclear', 'energy', 'temperature', 'find', 'qualitative', 'result', 'change', 'different', 'nuclear', 'physic', 'input', 'property', 'explosion', 'dynamic', 'magnetic', 'field', 'configuration', 'dominant', 'factor', 'determining', 'ejecta', 'composition']
```

Εικόνα 2.Δ.7 Lemmatization (αντί Stemming)

3. Ευρετήριο (Indexing)

3.α. Δημιουργία της ανεστραμμένης δομής δεδομένων ευρετηρίου

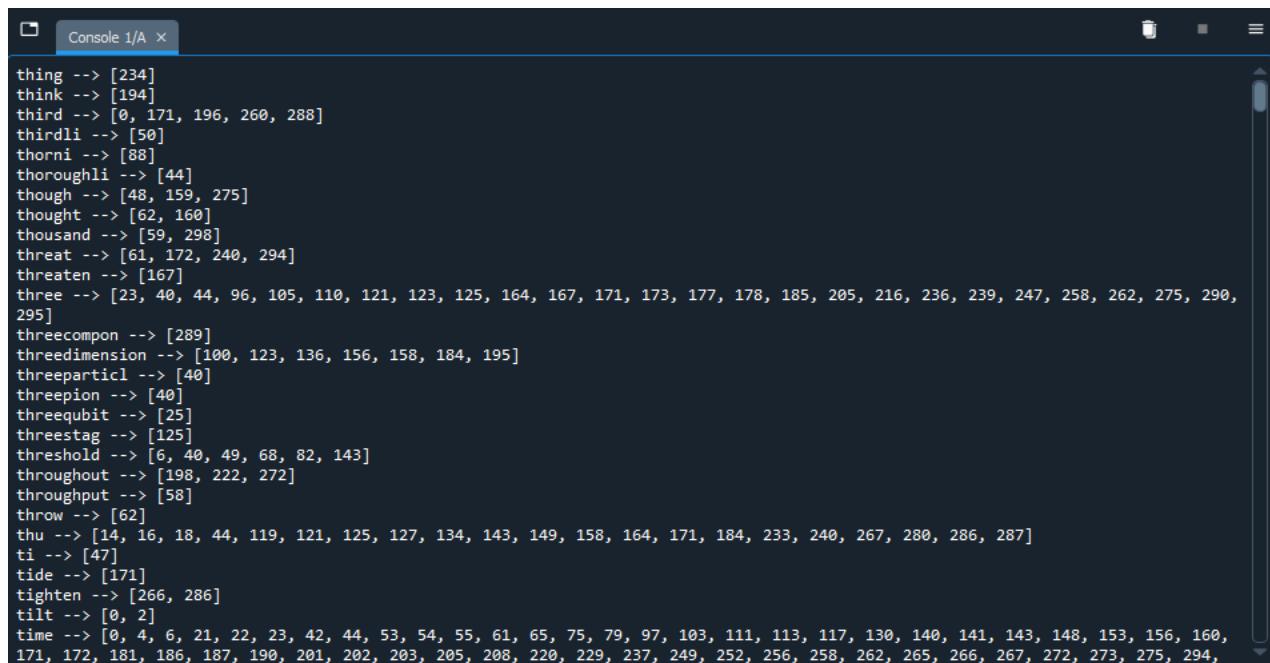
3.α.1 Σχεδιασμός

Ο σχεδιασμός της ανεστραμμένης δομής δεδομένων ευρετηρίου αποτελείται από την συλλογή των εργασιών για ευρετηρίαση ([βήμα 1](#)), την μετατροπή του κειμενικού περιεχομένου της κάθε εργασίας σε λίστα στοιχείων (tokenization) καθώς και την προεπεξεργασία τους ([βήμα 2](#)). Τέλος, αποτελείται από την ευρετηρίαση των εργασιών που περιέχουν τους όρους, δηλαδή, αντιστοιχεί τους όρους (tokens) που έχουν συλλεχθεί με τα συγκεκριμένα κείμενα στα οποία έχουν βρεθεί. Ουσιαστικά κάθε λέξη αντιστοιχεί σε έναν αριθμό ο οποίος με τη σειρά του δείχνει το κείμενο που βρίσκεται η λέξη. Για παράδειγμα, εάν μία λέξη που αντιστοιχεί στα νούμερα 1,3,7 ξέρουμε ότι βρίσκεται στα κείμενα με doc_id 1,3,7.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

3.a.2 Υλοποίηση

Για την υλοποίηση της ανεστραμμένης δομής δεδομένων ευρετηρίου χρειάστηκε η αρχικοποίηση ενός κενού ανεστραμμένου ευρετηρίου, έπειτα ο διαχωρισμός των λέξεων από τις περιλήψεις για κάθε περίληψη ξεχωριστά σε λεκτικές μονάδες (tokenization). Στη συνέχεια, προστίθονται όλες οι προεπεξεργασμένες λέξεις στο ευρετήριο, ενώ αν κάποια λέξη υπάρχει ήδη εντός αυτού απλά προστίθεται στη λέξη κλειδί του ευρετηρίου το id της εργασίας που έχει εντοπιστεί η λέξη, αποφεύγοντας έτσι τις διπλότυπες έννοιες εντός του ευρετηρίου. Τέλος, πραγματοποιείται ταξινόμηση των κλειδιών του λεξικού και των στοιχείων του.



```
Console 1/A ×
thing --> [234]
think --> [194]
third --> [0, 171, 196, 260, 288]
thirdli --> [50]
thorni --> [88]
thoroughli --> [44]
though --> [48, 159, 275]
thought --> [62, 160]
thousand --> [59, 298]
threat --> [61, 172, 240, 294]
threaten --> [167]
three --> [23, 40, 44, 96, 105, 110, 121, 123, 125, 164, 167, 171, 173, 177, 178, 185, 205, 216, 236, 239, 247, 258, 262, 275, 290, 295]
threecompon --> [289]
threedimension --> [100, 123, 136, 156, 158, 184, 195]
threeparticl --> [40]
threepion --> [40]
threequbit --> [25]
threestag --> [125]
threshold --> [6, 40, 49, 68, 82, 143]
throughout --> [198, 222, 272]
throughput --> [58]
throw --> [62]
thu --> [14, 16, 18, 44, 119, 121, 125, 127, 134, 143, 149, 158, 164, 171, 184, 233, 240, 267, 280, 286, 287]
ti --> [47]
tide --> [171]
tighten --> [266, 286]
tilt --> [0, 2]
time --> [0, 4, 6, 21, 22, 23, 42, 44, 53, 54, 55, 61, 65, 75, 79, 97, 103, 111, 113, 117, 130, 140, 141, 143, 148, 153, 156, 160, 171, 172, 181, 186, 187, 190, 201, 202, 203, 205, 208, 220, 229, 237, 249, 252, 256, 258, 262, 265, 266, 267, 272, 273, 275, 294,
```

Εικόνα 3.A.1 Η ανεστραμμένη δομή δεδομένων ευρετηρίου

3.a.3 Αξιολόγηση

Η ανεστραμμένη δομή δεδομένων ευρετηρίου αποδίδει σημαντικά στην αναζήτηση όρων σ' ένα κείμενο και γι' αυτό αποτελεί ισχυρό εργαλείο για τις μηχανές αναζήτησης. Η αντιστοίχιση των όρων με τα κείμενα στα οποία εμφανίζεται, δημιουργεί μία σημαντική εικόνα για την συγχόνηση των όρων στη συλλογή με τις εργασίες, γεγονός που ενισχύει το πόσο σημαντική πληροφορία είναι ένας όρος για την συλλογή. Η βέλτιστη ευρετηρίαση από άποψης αποτελεσματικότητας, θα ήταν να αναγράφεται και η/οι θέση/εις του όρου μέσα στο κείμενο, ώστε οι αναζητήσεις να είναι πιο αποτελεσματικές και ακριβείς, καθώς με αυτό τον τρόπο παρέχεται περισσότερη πληροφορία στον χρήστη. Ωστόσο, αυτή η μεθοδολογία δεν επιλέχθηκε κατά την υλοποίηση του ευρετηρίου της τοπικής μηχανής αναζήτησης, καθώς, το αποθετήριο είναι μικρό και δεν χρήζει ανάγκη μίας τέτοιας υλοποίησης που ξοδεύει χώρο στην μνήμη και χρόνο στην αναζήτηση λόγω του μεγάλου μεγέθους της.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

3.β. Αποθήκευση του ευρετηρίου σε μία δομή δεδομένων

3.β.1 Σχεδιασμός

Η δημιουργία της ανεστραμμένης δομής δεδομένων ευρετηρίου έχει ολοκληρωθεί στο [βήμα 3.α](#), οπότε αυτό που απομένει είναι να αποθηκευτεί σε μία δομή δεδομένων της γλώσσας Python, η οποία θα είναι αποδοτική ως προς την αναζήτηση όρων. Το σχεδιαστικό μοντέλο καταλήγει στην επιλογή της δομής λεξικού για την αποθήκευση του ευρετηρίου.

3.β.2 Υλοποίηση

Η δομή του λεξικού περιλαμβάνει δύο πεδία, τα κλειδιά (keys) και τα περιεχόμενα (values). Στην υλοποίηση της τοπικής μηχανής αναζήτησης, οι λέξεις-κλειδιά είναι όλοι οι μοναδικοί όροι της συλλογής εργασιών (terms), ενώ περιεχόμενα είναι λίστες (posting lists) με τους μοναδικούς ακέραιους αριθμούς (doc_id) που δηλώνουν τα κείμενα μέσα στα οποία εμφανίζεται ο όρος-κλειδί.

3.β.3 Αξιολόγηση

Η επιλογή της δομής λεξικού για την αποθήκευση του ευρετηρίου αξιολογείται από το γεγονός ότι αποδίδει αποτελεσματικότητα στην αναζήτηση χάρις στην αντιστοίχιση των όρων-κλειδιών με τις λίστες με τα κείμενα στα οποία εμφανίζονται. Ωστόσο, ενδέχεται να υπάρχει σημαντική κατανάλωση μνήμης λόγω του μεγάλου μεγέθους του ευρετηρίου που προκύπτει από το πλήθος των μοναδικών όρων-κλειδιών. Παρόλα αυτά, οι απαιτήσεις της τοπικής μηχανής αναζήτησης εμβαθύνουν στην ακρίβεια και την αποτελεσματικότητα της αναζήτησης και όχι στην εξοικονόμηση χώρου. Με επιλογή μίας άλλης δομής (π.χ. πίνακας, λίστα, πλειάδα κλπ) θα υπήρχε σημαντική βελτίωση στον χώρο, αλλά θα υπήρχε σημαντικό αντίκτυπο στο αποτέλεσμα της αναζήτησης. Κλείνοντας, το αποθετήριο είναι μικρό οπότε η κατανάλωση μνήμης δεν θα ήταν τόση ώστε να προκαλέσει σημαντικό πρόβλημα στην απόδοση της μηχανής αναζήτησης.

4. Μηχανή αναζήτησης (Search Engine)

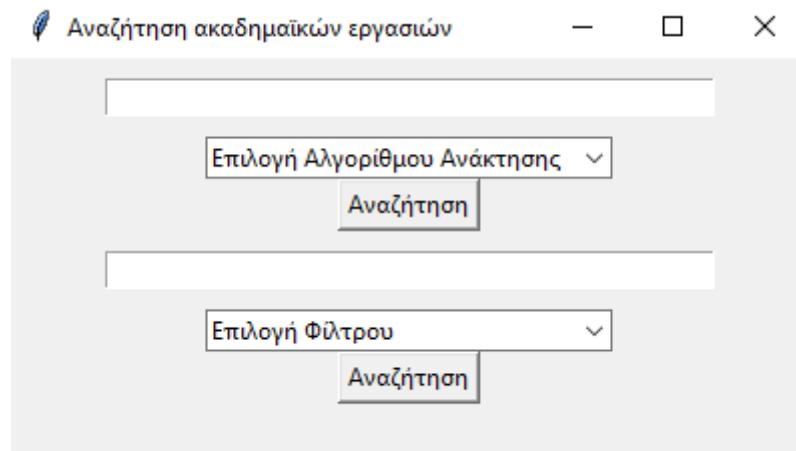
4.α. Ανάπτυξη διεπαφής χρήστη για αναζήτηση εργασιών

4.α.1 Σχεδιασμός

Προκειμένου η μηχανή αναζήτησης να είναι πιο φιλική στον χρήστη σχεδιάστηκε μια διεπαφή χρήστη (GUI) για την αναζήτηση των εργασιών. Πιο συγκεκριμένα η διεπαφή αυτή περιέχει:

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Παράθυρο διεπαφής χρήστη
- Πεδίο εισαγωγής ερωτήματος χρήστη
- Πεδίο επιλογής αλγορίθμου ανάκτησης (Boolean Retrieval, Vector Space Model, Probabilistic Retrieval Model)
- Κουμπί αναζήτησης αποτελεσμάτων αλγόριθμου ανάκτησης
- Πεδίο εισαγωγής κριτηρίου φίλτρου
- Πεδίο επιλογής φίλτρου
- Κουμπί αναζήτησης αποτελεσμάτων φίλτρου



Εικόνα 4.A.1 Η διεπαφή χρήστη

4.a.2 Υλοποίηση

Η υλοποίηση της διεπαφής χρήστη, ορίστηκε σε ξεχωριστό module (search_engine.py) και συγκεκριμένα χρησιμοποιήθηκε η βιβλιοθήκη tkinter. Η διεπαφή εξυπηρετεί τις δύο βασικές λειτουργίες της αναζήτησης εργασιών με βάση ένα ερώτημα χρήστη και φιλτράρισμα αποτελεσμάτων με βάση συγκεκριμένα κριτήρια, όπως είναι ο συγγραφέας ή ημερομηνία δημοσίευσης. Η λειτουργία της διεπαφής και της αναζήτησης εργασιών υλοποιούνται σε μία κλάση με ιδιότητες το αποθετήριο με προεπεξεργασμένα τα πεδία «abstract» ([βήμα 2](#)) και το ανεστραμμένο ευρετήριο ([βήμα 3](#)).

Οι υλοποιήσεις των εξαρτημάτων της διεπαφής που αναφέρθηκαν στον [σχεδιασμό](#) είναι οι εξής:

- Πεδίο εισαγωγής ερωτήματος χρήστη: Ο χρήστης εισάγει το ερώτημα
- Πεδίο επιλογής αλγορίθμου ανάκτησης: Ο χρήστης επιλέγει από ένα menu έναν από τους 3 αλγόριθμους ανάκτησης (Boolean Retrieval, Vector Space Model, Probabilistic Retrieval Model), στον οποίο θέλει να εφαρμοστεί για την ανάκτηση εργασιών
- Κουμπί αναζήτησης αποτελεσμάτων αλγόριθμου ανάκτησης: Με το πάτημα του κουμπιού «Αναζήτηση» εμφανίζονται τα αποτελέσματα του ερωτήματος στο τερματικό παράθυρο, με βάση τον αλγόριθμο ανάκτησης που επέλεξε ο χρήστης
- Πεδίο επιλογής κριτηρίου φίλτρου: Ο χρήστης επιλέγει από ένα menu (εφόσον επιθυμεί) ένα από τα 2 κριτήρια φίλτραρισμάτων (Συγγραφέας, Ημερομηνία δημοσίευσης)
- Πεδίο εισαγωγής κριτηρίου φίλτρου: Ο χρήστης εισάγει το φίλτρο (όνομα συγγραφέα ή ημερομηνία δημοσίευσης)

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Κουμπί αναζήτησης αποτελεσμάτων φίλτρου: Με το πάτημα του κουμπιού «Αναζήτηση» εμφανίζονται τα αποτελέσματα του φίλτραρισματος στο τερματικό παράθυρο, με βάση το κριτήριο φίλτραρισματος που επέλεξε ο χρήστης

4.α.3 Αξιολόγηση

Η διεπαφή αν και απλή, πληροί τα κριτήρια ώστε να μπορεί ο χρήστης με ευκολία να αναζητήσει εργασίες βάση ορισμένων λέξεων που θέτει ως ερώτημα αναζήτησης. Εξίσου εύκολα επιτρέπει το φίλτραρισμα με βάση συγκεκριμένα κριτήρια (ημερομηνία δημοσίευσης, συγγραφέα), συνεπώς δεν υπάρχει κάτι άλλο στη διεπαφή το οποίο δεν υπάρχει ως λειτουργία προς τον χρήστη. Η διεπαφή φροντίζει για την σωστή είσοδο στοιχείων στην μηχανή αναζήτησης, για την επεξεργασία στοιχείων και για κάποιο ενδεχόμενο φίλτραρισμα. Η χρησιμότητα της κλάσης αυτής είναι καταλυτική, μιας και με το κλείσιμο της διεπαφής τερματίζει και το πρόγραμμα. Ίσως, θα ήταν πιο προτιμητέο να μην υπήρχαν τόσες συναρτήσεις συγκεντρωμένες στην κλάση της διεπαφής καθαρά ως θέμα ευελιξίας και αναγνωσιμότητας του κώδικα.

4.β. Υλοποίηση αλγορίθμων ανάκτησης

4.β.1 Σχεδιασμός

Boolean Retrieval

Ο αλγόριθμος Boolean Retrieval που εφαρμόζεται στη μηχανή αναζήτησης, βασίζεται στην άλγεβρα Boole και τη θεωρία συνόλων, χρησιμοποιώντας τρεις βασικές λέξεις-κλειδιά: OR, AND και NOT. Ακολουθεί την διαδικασία της Επεξεργασίας ερωτήματος (Query Processing). Αυτές οι λειτουργίες επηρεάζουν τα αποτελέσματα της αναζήτησης ανάλογα με τη θέση τους στο ερώτημα.

- Η λέξη-κλειδί NOT, όταν τοποθετείται πριν από μια λέξη, αναζητά τα κείμενα που δεν περιέχουν την εν λόγω λέξη.
- Η λέξη-κλειδί AND, όταν τοποθετείται μεταξύ δύο λέξεων, αναζητά τα κείμενα που περιέχουν και τις δύο λέξεις.
- Η λέξη-κλειδί OR, όταν τοποθετείται μεταξύ δύο λέξεων, αναζητά τα κείμενα που περιέχουν οποιαδήποτε από τις δύο λέξεις.

Επιπλέον, ο αλγόριθμος λαμβάνει υπόψη την προτεραιότητα των πράξεων σε σύνθετα Boolean ερωτήματα χρήστη, εκτελώντας τις πράξεις με αυτή την σειρά προτεραιότητας:

1. Πράξεις με τους λογικούς τελεστές NOT μέσα σε παρένθεση (από αριστερά προς τα δεξιά)

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

2. Πράξεις με τους τελεστές AND ή OR μέσα σε παρένθεση (από αριστερά προς τα δεξιά)
3. Πράξεις με τους λογικούς τελεστές NOT (από αριστερά προς τα δεξιά)
4. Πράξεις με τους τελεστές AND ή OR (από αριστερά προς τα δεξιά)

Με αυτόν τον τρόπο, ο αλγόριθμος μπορεί να προσαρμοστεί για να ανταποκριθεί σε πολύπλοκα ερωτήματα αναζήτησης.

Vector Space Model

Ο αλγόριθμος Vector Space Model που εφαρμόζεται στη μηχανή αναζήτησης, αναπαριστά τα κείμενα και τα ερωτήματα χρηστών ως διανύσματα σ' έναν n-διάστατο χώρο. Από τους αλγόριθμους κατάταξης ακολουθεί τους TF-IDF και Cosine Similarity για τον υπολογισμό της ομοιότητας μεταξύ του ερωτήματος χρήστη και των κειμένων της συλλογής. Λόγω του προβλήματος με τις μεγάλες Ευκλείδεις αποστάσεις των διανυσμάτων, ο υπολογισμός του συνημιτόνου της γωνίας που σχηματίζει το διάνυσμα του ερωτήματος με το διάνυσμα κάποιου κειμένου της συλλογής, είναι πιο αποδοτικός. Όσο πιο κοντά είναι τα διανύσματα, τόσο πιο πλησιέστερο είναι το ερώτημα με το κείμενο, δηλαδή, το συνημίτονο της γωνίας τείνει προς το 1, όπου 1 σημαίνει ταυτόσημα.

Probabilistic Retrieval Model

Ο αλγόριθμος Probabilistic Retrieval Model που εφαρμόζεται στη μηχανή αναζήτησης, αναζητά εργασίες με βάση το μοντέλο των πιθανοτήτων. Ουσιαστικά υπολογίζει την πιθανότητα ένα κείμενο να είναι σχετικό με το ερώτημα χρήστη. Από τους αλγόριθμους κατάταξης ακολουθεί τους TF-IDF και Okapi BM25 για τον υπολογισμό των πιθανοτήτων.

4.β.2 Υλοποίηση

Boolean Retrieval

Η υλοποίηση του Boolean Retrieval πραγματοποιήθηκε στο ίδιο module (search_engine.py) με την διεπαφή χρήστη και συγκεκριμένα στην ίδια κλάση. Σαν είσοδο, ο αλγόριθμος παίρνει τα εξής δεδομένα:

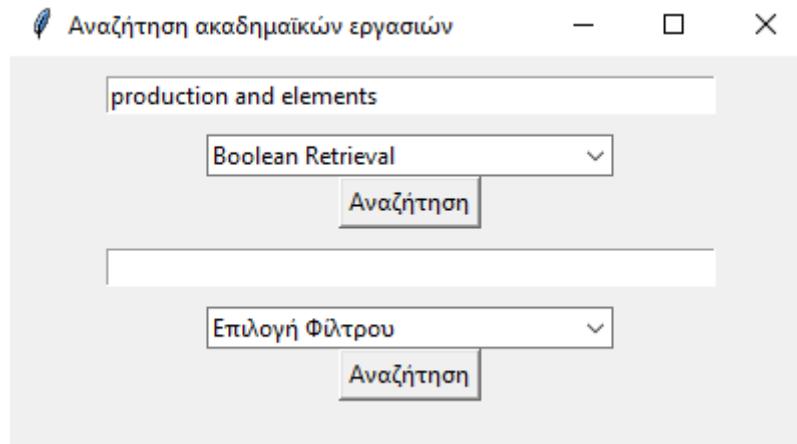
- Το Boolean ερώτημα που θα δώσει ο χρήστης για να κάνει την αναζήτηση
- Το ανεστραμμένο ευρετήριο που θα κάνει την απαραίτητη αντιστοίχιση μεταξύ των λέξεων και των κειμένων

Η λειτουργία του αλγορίθμου είναι στην δημιουργία μίας στοίβας από απλά Boolean υπο-ερωτήματα του σύνθετου ή και απλού Boolean ερωτήματος χρήστη. Η θέση του κάθε υπο-ερωτήματος στην στοίβα καθορίζεται από την σειρά προτεραιότητα των πράξεων που αναφέρθηκαν στον σχεδιασμό. Για κάθε υπο-ερώτημα εκτελείται η πράξη από την διαδικασία της επεξεργασίας ερωτήματος. Πρώτο μέλημα του αλγόριθμου είναι αν υπάρχουν παρενθέσεις. Στην περίπτωση που υπάρχουν παρενθέσεις, ο αλγόριθμος εκχωρεί στην στοίβα πρώτα τις λογικές πράξεις εντός των παρενθέσεων και έπειτα όποιες άλλες πράξεις μένουν. Αν δεν υπάρχουν παρενθέσεις στο query του χρήστη, η αναζήτηση γίνεται κανονικά από αριστερά προς τα δεξιά.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Η κάθε λέξη του query πρέπει να εξεταστεί σωστά για να ελεγχθεί αν αντιστοιχεί σε κάποια λέξη-κλειδί (and, or, not) ή λέξη που βρίσκεται εντός του ευρετηρίου. Γι' αυτό το λόγο, οι όροι του query πρέπει να γίνουν tokens, μέσω των οποίων θα μπορέσουν να εντοπιστούν ενδεχόμενες λέξεις-κλειδιά που θα επιτρέψουν την σωστή αντιστοίχιση μεταξύ λέξεων του query και λέξεις που βρίσκονται στο ευρετήριο.

Μερικές ενδεικτικές μελέτες περιπτώσεων από ερωτήματα χρηστών είναι οι εξής:



```
In [17]: runfile('C:/Users/billa/source/repos/SearchEngine/main.py', wdir='C:/Users/billa/source/repos/SearchEngine')
Reloaded modules: web_crawler, text_preprocessing, inverted_index, query_processing, ranking, search_engine

===== Ερώτημα αναζήτησης : production and elements =====
===== Αλγόριθμος ανάκτησης : Boolean Retrieval =====
-----
#1
-----
Document ID : 0
Title      : Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors    : M. Reichert, M. Bugli, J. Guilet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Subjects   : High Energy Astrophysical Phenomena
Abstract   : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $\beta^{\{-\}}$-decays and $\beta^{\{-\}}$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
Comments   :
Date       : 25 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14402
```

Εικόνα 4.B.1 Απλό Boolean query με AND

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Αναζήτηση ακαδημαϊκών εργασιών

production or elements

Boolean Retrieval

Αναζήτηση

Επιλογή Φίλτρου

Αναζήτηση

```
===== Ερώτημα αναζήτησης : production or elements =====
===== Αλγόριθμος ανάκτησης : Boolean Retrieval =====
#1
Document ID : 0
Title : Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors : M. Reichert, M. Bugli, J. Guilet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Subjects : High Energy Astrophysical Phenomena
Abstract : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $\beta^{+}{}$-decays and $\beta^{0}{}$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
Comments :
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14402
#2
Document ID : 42
Title : How far can we see back in time in high-energy collisions using charm quarks?
Authors : Laszlo Gyulai, Gabor Biro, Robert Vertesi, Gergely Gabor Barnafoldi
Subjects : High Energy Physics - Phenomenology, Nuclear Theory
Abstract : We use open charm production to estimate how far we can see back in time in high-energy hadron-hadron collisions. We analyze the transverse momentum distributions of the identified D mesons from pp, p-Pb and A-A collisions at the ALICE and STAR experiments covering the energy range from  $\sqrt{s_{\text{NN}}} = 200$  GeV up to 7 TeV. Within a non-extensive statistical framework, the common Tsallis parameters for D mesons represent higher temperature and more degrees of freedom than that of light-flavour hadrons. The production of D mesons corresponds to a significantly earlier proper time,  $\tau_D = (0.18 \pm 0.06) \tau_{\text{LF}}$ .
Comments : 18 pages, 6 figures, 1 table
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14282
#3
Document ID : 45
Title : Phenomenology of TMD parton distributions in Drell-Yan and $Z^0$ boson production in a hadron structure oriented
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

taken from the 2014 Global Adult Population Survey or the Global Entrepreneurship Monitor project. The propensity for innovation amongst tourism entrepreneurs has a statistically significant relationship to gender, age, level of education and informal investments in previous businesses.

Comments : Journal ref: Sustainability 12:5003 (2020)
Date : 5 December 2023
PDF_URL : <https://arxiv.org/pdf/2401.13679>

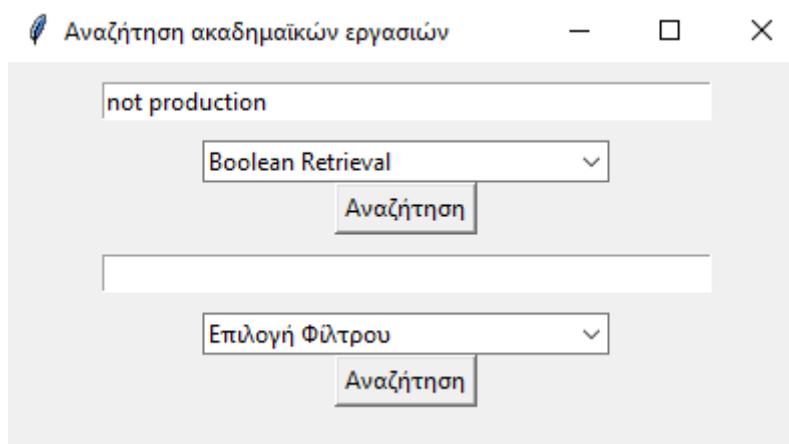
#19

Document ID : 236
Title : Realized Stochastic Volatility Model with Skew-t Distributions for Improved Volatility and Quantile Forecasting
Authors : Makoto Takahashi, Yuta Yamauchi, Toshiaki Watanabe, Yasuhiro Omori
Subjects : Econometrics
Abstract : Forecasting volatility and quantiles of financial returns is essential for accurately measuring financial tail risks, such as value-at-risk and expected shortfall. The critical elements in these forecasts involve understanding the distribution of financial returns and accurately estimating volatility. This paper introduces an advancement to the traditional stochastic volatility model, termed the realized stochastic volatility model, which integrates realized volatility as a precise estimator of volatility. To capture the well-known characteristics of return distribution, namely skewness and heavy tails, we incorporate three types of skew-t distributions. Among these, two distributions include the skew-normal feature, offering enhanced flexibility in modeling the return distribution. We employ a Bayesian estimation approach using the Markov chain Monte Carlo method and apply it to major stock indices. Our empirical analysis, utilizing data from US and Japanese stock indices, indicates that the inclusion of both skewness and heavy tails in daily returns significantly improves the accuracy of volatility and quantile forecasts.
Comments :
Date : 23 January 2024
PDF_URL : <https://arxiv.org/pdf/2401.13179>

#20

Document ID : 248
Title : Optimal design of a local renewable electricity supply system for power-intensive production processes with demand response
Authors : Sonja H. M. Germescheid, Benedikt Nilges, Niklas von der Assen, Alexander Mitsos, Manuel Dahmen
Subjects : Optimization and Control
Abstract : This work studies synergies arising from combining industrial demand response and local renewable electricity supply. To this end, we optimize the design of a local electricity generation and storage system with an integrated demand response scheduling of a continuous power-intensive production process in a multi-stage problem. We optimize both total annualized cost and global warming impact and consider local photovoltaic and wind electricity generation, an electric battery, and electricity trading on day-ahead and intraday market. We find that installing a battery can reduce emissions and enable large trading volumes on the electricity markets, but significantly increases cost. Economic and ecologic process and battery operation are driven primarily by the electricity price and grid emission factor, respectively, rather than locally generated electricity. A parameter study reveals that economic savings from the local system and flexibilizing the process behave almost additive.
Comments : manuscript (32 pages, 9 figures, 6 tables), supporting materials (11 pages, 9 figures, 2 tables)
Date : 23 January 2024
PDF_URL : <https://arxiv.org/pdf/2401.12759>

Εικόνα 4.B.2 Απλό Boolean query με OR



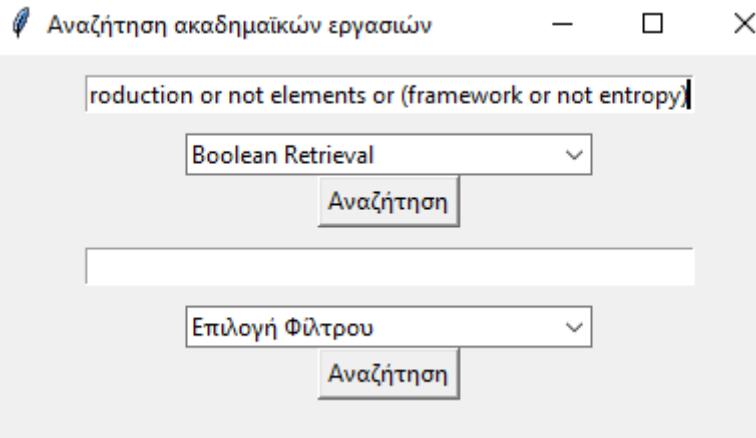
ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
===== Ερύτημα ανάζητησης : not production =====
===== Αλγόριθμος ανάζησης : Boolean Retrieval =====

#1
-----
Document ID : 1
Title : pix2gestalt: Amodal Segmentation by Synthesizing Wholes
Authors : Ege Ozguroglu, Ruoshi Liu, Didac Suris, Dian Chen, Achal Dave, Pavel Tokmakov, Carl Vondrick
Subjects : Computer Vision and Pattern Recognition, Machine Learning
Abstract : We introduce pix2gestalt, a framework for zero-shot amodal segmentation, which learns to estimate the shape and appearance of whole objects that are only partially visible behind occlusions. By capitalizing on large-scale diffusion models and transferring their representations to this task, we learn a conditional diffusion model for reconstructing whole objects in challenging zero-shot cases, including examples that break natural and physical priors, such as art. As training data, we use a synthetically curated dataset containing occluded objects paired with their whole counterparts. Experiments show that our approach outperforms supervised baselines on established benchmarks. Our model can furthermore be used to significantly improve the performance of existing object recognition and 3D reconstruction methods in the presence of occlusions.
Comments : Website: https://gestalt.cs.columbia.edu/
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14398.pdf
-----
#2
-----
Document ID : 2
Title : Entanglement entropy and deconfined criticality: emergent  $SO(5)$  symmetry and proper lattice bipartition
Authors : Jonathan D'Emidio, Anders W. Sandvik
Subjects : Strongly Correlated Electrons, High Energy Physics - Lattice
Abstract : We study the Rényi entanglement entropy (EE) of the two-dimensional  $J=Q$  model, the emblematic quantum spin model of deconfined criticality at the phase transition between antiferromagnetic and valence-bond-solid ground states. Quantum Monte Carlo simulations with an improved EE scheme reveal critical corner contributions that scale logarithmically with the system size, with a coefficient in remarkable agreement with the form expected from a large- $N$  conformal field theory with  $SO(N=5)$  symmetry. However, details of the bipartition of the lattice are crucial in order to observe this behavior. If the subsystem for the reduced density matrix does not properly accommodate valence-bond fluctuations, logarithmic contributions appear even for corner-less bipartitions. We here use a  $45^\circ$  tilted cut on the square lattice. Beyond supporting an  $SO(5)$  deconfined quantum critical point, our results for both the regular and tilted cuts demonstrate important microscopic aspects of the EE that are not captured by conformal field theory.
Comments : 5 pages, 3 figures
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14396.pdf
-----
#3
-----
Document ID : 3
Title : Summing up perturbation series around superintegrable point
Authors : A. Mironov, A. Morozov, A. Popolitov, Sh. Shakirov
Subjects : High Energy Physics - Theory, Mathematical Physics
Abstract : We work out explicit formulas for correlators in the Gaussian matrix model perturbed by a logarithmic potential, i.e. by inserting Miwa variables. In this paper, we concentrate on the example of a single Miwa variable. The ordinary Gaussian
-----
#19
-----
Document ID : 19
Title : Uncovering Heterogeneity of Solar Flare Mechanism With Mixture Models
Authors : Bach Viet Do, Yang Chen, XuanLong Nguyen, Ward Manchester
Subjects : Solar and Stellar Astrophysics, Applications, Methodology
Abstract : The physics of solar flares occurring on the Sun is highly complex and far from fully understood. However, observations show that solar eruptions are associated with the intense kilogauss fields of active regions, where free energies are stored with field-aligned electric currents. With the advent of high-quality data sources such as the Geostationary Operational Environmental Satellites (GOES) and Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI), recent works on solar flare forecasting have been focusing on data-driven methods. In particular, black box machine learning and deep learning models are increasingly adopted in which underlying data structures are not modeled explicitly. If the active regions indeed follow the same laws of physics, there should be similar patterns shared among them, reflected by the observations. Yet, these black box models currently used in the literature do not explicitly characterize the heterogeneous nature of the solar flare data, within and between active regions. In this paper, we propose two finite mixture models designed to capture the heterogeneous patterns of active regions and their associated solar flare events. With extensive numerical studies, we demonstrate the usefulness of our proposed method for both resolving the sample imbalance issue and modeling the heterogeneity for rare energetic solar flare events.
Comments :
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14345.pdf
-----
#20
-----
Document ID : 20
Title : From the Choi Formalism in Infinite Dimensions to Unique Decompositions of Generators of Completely Positive Dynamical Semigroups
Authors : Frederik vom Ende
Subjects : Functional Analysis, Mathematical Physics, Quantum Physics
Abstract : Given any separable complex Hilbert space, any trace-class operator  $B$  which does not have purely imaginary trace, and any generator  $L$  of a norm-continuous one-parameter semigroup of completely positive maps we prove that there exists a unique bounded operator  $K$  and a unique completely positive map  $\Phi$  such that (i)  $L = K(\cdot \dot{+} (\cdot \dot{+} K^*))$ , (ii) the superoperator  $\Phi(B^*(\cdot \dot{+} B))$  is trace class and has vanishing trace, and (iii)  $\|\text{tr } \Phi(B^*)\|$  is a real number. Central to our proof is a modified version of the Choi formalism which relates completely positive maps to positive semi-definite operators. We characterize when this correspondence is injective and surjective, respectively, which in turn explains why the proof idea of our main result cannot extend to non-separable Hilbert spaces. In particular, we find examples of positive semi-definite operators which have empty pre-image under the Choi formalism as soon as the underlying Hilbert space is infinite-dimensional.
Comments : 25+3 pages. Generalizes arXiv:2310.04037 to infinite dimensions. To be submitted to J. Funct. Anal
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14344.pdf
```

Ευκόνα 4.B.3 Απλό Boolean query με NOT

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



```
===== Ερώτημα αναζήτησης : production or not elements or (framework or not entropy) =====
===== Αλγόριθμος αναζήτησης : Boolean Retrieval =====
-----
#1
-----
Document ID : 0
Title : Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors : M. Reichert, M. Bugli, J. Guilet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Subjects : High Energy Astrophysical Phenomena
Abstract : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $\beta^{+}$-$\beta$-decays and $\beta^{+}$-$\beta$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
Comments :
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14402.pdf
-----
#2
-----
Document ID : 1
Title : pix2gestalt: Amodal Segmentation by Synthesizing Wholes
Authors : Ege Ozguroglu, Ruoshi Liu, Didac Suris, Dian Chen, Achal Dave, Pavel Tokmakov, Carl Vondrick
Subjects : Computer Vision and Pattern Recognition, Machine Learning
Abstract : We introduce pix2gestalt, a framework for zero-shot amodal segmentation, which learns to estimate the shape and appearance of whole objects that are only partially visible behind occlusions. By capitalizing on large-scale diffusion models and transferring their representations to this task, we learn a conditional diffusion model for reconstructing whole objects in challenging zero-shot cases, including examples that break natural and physical priors, such as art. As training data, we use a synthetically curated dataset containing occluded objects paired with their whole counterparts. Experiments show that our approach outperforms supervised baselines on established benchmarks. Our model can furthermore be used to significantly improve the performance of existing object recognition and 3D reconstruction methods in the presence of occlusions.
Comments : Website: https://gestalt.cs.columbia.edu/
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14398.pdf
-----
#3
-----
Document ID : 2
Title : Entanglement entropy and deconfined criticality: emergent $SO(5)$ symmetry and proper lattice binartition
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

the different patterns of Bethe roots of the reduced Bethe ansatz equations for the different boundary parameters. According to them, we obtain the densities of states, ground state energy density and surface energy. Our results show that the system has the stable boundary bound states when the boundary magnetic fields satisfy some constraints.

Comments : 13 pages, 3 figures
Date : 25 January 2024
PDF_URL : <https://arxiv.org/pdf/2401.14356>

#19

Document ID : 18
Title : Initial data for Minkowski stability with arbitrary decay
Authors : Allen Juntao Fang, Jérémie Szeftel, Arthur Touati
Subjects : Analysis of PDEs, General Relativity and Quantum Cosmology, Mathematical Physics
Abstract : We construct and parametrize solutions to the constraint equations of general relativity in a neighborhood of Minkowski spacetime with arbitrary prescribed decay properties at infinity. We thus provide a large class of initial data for the results on stability of Minkowski which include a mass term in the asymptotics. Due to the symmetries of Minkowski, a naive linear perturbation fails. Our construction is based on a simplified conformal method, a reduction to transverse traceless perturbations and a nonlinear fixed point argument where we face linear obstructions coming from the cokernels of both the linearized constraint operator and the Laplace operator. To tackle these obstructions, we introduce a well-chosen truncated black hole around which to perturb. The control of the parameters of the truncated black hole is the most technical part of the proof, since its center of mass and angular momentum could be arbitrarily large.
Comments : 86 pages
Date : 25 January 2024
PDF_URL : <https://arxiv.org/pdf/2401.14353>

#20

Document ID : 19
Title : Uncovering Heterogeneity of Solar Flare Mechanism With Mixture Models
Authors : Bach Viet Do, Yang Chen, XuanLong Nguyen, Ward Manchester
Subjects : Solar and Stellar Astrophysics, Applications, Methodology
Abstract : The physics of solar flares occurring on the Sun is highly complex and far from fully understood. However, observations show that solar eruptions are associated with the intense kilogauss fields of active regions, where free energies are stored with field-aligned electric currents. With the advent of high-quality data sources such as the Geostationary Operational Environmental Satellites (GOES) and Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI), recent works on solar flare forecasting have been focusing on data-driven methods. In particular, black box machine learning and deep learning models are increasingly adopted in which underlying data structures are not modeled explicitly. If the active regions indeed follow the same laws of physics, there should be similar patterns shared among them, reflected by the observations. Yet, these black box models currently used in the literature do not explicitly characterize the heterogeneous nature of the solar flare data, within and between active regions. In this paper, we propose two finite mixture models designed to capture the heterogeneous patterns of active regions and their associated solar flare events. With extensive numerical studies, we demonstrate the usefulness of our proposed method for both resolving the sample imbalance issue and modeling the heterogeneity for rare energetic solar flare events.
Comments :
Date : 25 January 2024
PDF_URL : <https://arxiv.org/pdf/2401.14345>

Εικόνα 4.B.4 Σύνθετο Boolean query

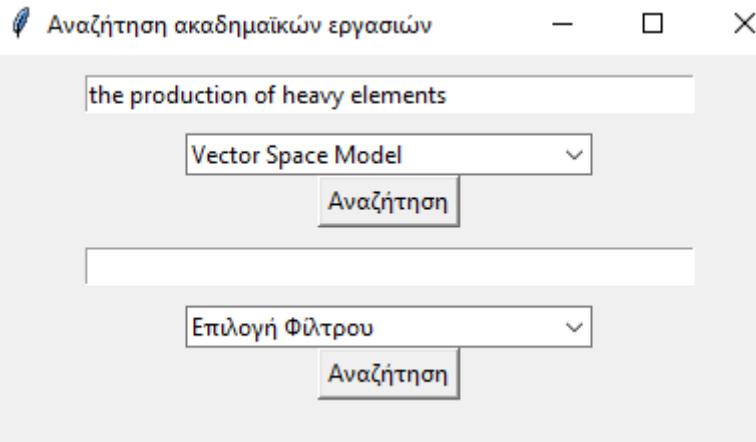
Vector Space Model

Η υλοποίηση του Vector Space Model πραγματοποιήθηκε στο ίδιο module (search_engine.py) με την διεπαφή χρήστη και συγκεκριμένα στην ίδια κλάση. Σαν είσοδο, ο αλγόριθμος παίρνει το ερώτημα χρήστη και οι υλοποιήσεις που λαμβάνουν χώρα είναι οι εξής:

- Προεπεξεργασία κειμένου: [Βίμα 2](#)
- Υπολογισμός TF-IDF των όρων των κειμένων: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Υπολογισμός TF-IDF των όρων του ερωτήματος: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Υπολογισμός των συνημιτόνων μεταξύ των κειμένων και του ερωτήματος: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Κατάταξη αποτελεσμάτων (Ranking)

Μερικές ενδεικτικές μελέτες περιπτώσεων από ερωτήματα χρηστών είναι οι εξής:

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



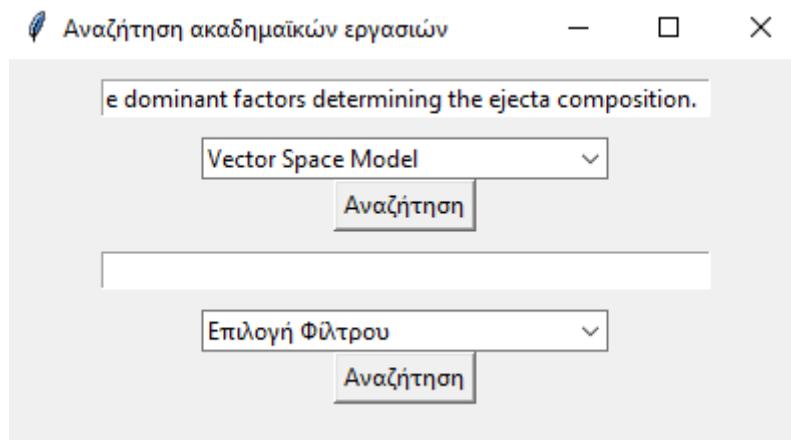
```
Console 1/A ×

===== Ερύτημα αναζήτησης : the production of heavy elements =====
===== Αλγόριθμος ανάκτησης : Vector Space Model =====
#1 Cosine Similarity: 0.1566
-----
Document ID : 0
Title      : Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors    : M. Reichert, M. Bugli, J. Guilet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Subjects   : High Energy Astrophysical Phenomena
Abstract   : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $β^{-}$$-decays and $β^{-}$$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
Comments   :
Date       : 25 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14402
#2 Cosine Similarity: 0.1539
-----
Document ID : 236
Title      : Realized Stochastic Volatility Model with Skew-t Distributions for Improved Volatility and Quantile Forecasting
Authors    : Makoto Takahashi, Yuta Yamauchi, Toshiaki Watanabe, Yasuhiro Omori
Subjects   : Econometrics
Abstract   : Forecasting volatility and quantiles of financial returns is essential for accurately measuring financial tail risks, such as value-at-risk and expected shortfall. The critical elements in these forecasts involve understanding the distribution of financial returns and accurately estimating volatility. This paper introduces an advancement to the traditional stochastic volatility model, termed the realized stochastic volatility model, which integrates realized volatility as a precise estimator of volatility. To capture the well-known characteristics of return distribution, namely skewness and heavy tails, we incorporate three types of skew-t distributions. Among these, two distributions include the skew-normal feature, offering enhanced flexibility in modeling the return distribution. We employ a Bayesian estimation approach using the Markov chain Monte Carlo method and apply it to major stock indices. Our empirical analysis, utilizing data from US and Japanese stock indices, indicates that the inclusion of both skewness and heavy tails in daily returns significantly improves the accuracy of volatility and quantile forecasts.
Comments   :
Date       : 23 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.13179
#3 Cosine Similarity: 0.0002
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
Console 1/A X
stepwise statistical multi-regression model with leave-one-out cross-validation. Under specific management conditions (e.g., three annual cuts) and from one to five months in advance, the generated model successfully provided a p-value<0.01 in correlation (t-test), a root mean square error percentage (%RMSE) of 14.6% and a 71.43% hit rate predicting above/below average years in terms of forage yield collection.
Comments : Journal ref: Gomara I, Bellocchi G, Martin R, Rodriguez-Fonseca B, Ruiz-Ramos M (2020) Agricultural and Forest Meteorology, 280, 107768
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14053
-----
#19 Cosine Similarity: 0.0262
-----
Document ID : 165
Title : FIMBA: Evaluating the Robustness of AI in Genomics via Feature Importance Adversarial Attacks
Authors : Hеorhi Skovorodnikov, Hoda Alkhzaimi
Subjects : Machine Learning, Cryptography and Security, Genomics
Abstract : With the steady rise of the use of AI in bio-technical applications and the widespread adoption of genomics sequencing, an increasing amount of AI-based algorithms and tools is entering the research and production stage affecting critical decision-making streams like drug discovery and clinical outcomes. This paper demonstrates the vulnerability of AI models often utilized downstream tasks on recognized public genomics datasets. We undermine model robustness by deploying an attack that focuses on input transformation while mimicking the real data and confusing the model decision-making, ultimately yielding a pronounced deterioration in model performance. Further, we enhance our approach by generating poisoned data using a variational autoencoder-based model. Our empirical findings unequivocally demonstrate a decline in model performance, underscored by diminished accuracy and an upswing in false positives and false negatives. Furthermore, we analyze the resulting adversarial samples via spectral analysis yielding conclusions for countermeasures against such attacks.
Comments : 15 pages, core code available at: https://github.com/HеorhiS/fimba-attack
Date : 19 January 2024
PDF_URL : https://arxiv.org/pdf/2401.10657
-----
#20 Cosine Similarity: 0.0229
-----
Document ID : 248
Title : Optimal design of a local renewable electricity supply system for power-intensive production processes with demand response
Authors : Sonja H. M. Germscheid, Benedikt Nilges, Niklas von der Assen, Alexander Mitsos, Manuel Dahmen
Subjects : Optimization and Control
Abstract : This work studies synergies arising from combining industrial demand response and local renewable electricity supply. To this end, we optimize the design of a local electricity generation and storage system with an integrated demand response scheduling of a continuous power-intensive production process in a multi-stage problem. We optimize both total annualized cost and global warming impact and consider local photovoltaic and wind electricity generation, an electric battery, and electricity trading on day-ahead and intraday market. We find that installing a battery can reduce emissions and enable large trading volumes on the electricity markets, but significantly increases cost. Economic and ecologic process and battery operation are driven primarily by the electricity price and grid emission factor, respectively, rather than locally generated electricity. A parameter study reveals that economic savings from the local system and flexibilizing the process behave almost additive.
Comments : manuscript (32 pages, 9 figures, 6 tables), supporting materials (11 pages, 9 figures, 2 tables)
Date : 23 January 2024
PDF_URL : https://arxiv.org/pdf/2401.12759
```

Εικόνα 4.B.5 Μικρό ερώτημα χρήστη (VSM)



ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
Console 1/A ×
=====
Ερώτημα αναζήτησης : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $\beta^{\{-\}}$-decays and $\beta^{\{+\}}$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
=====
Αλγόριθμος ανάκτησης : Vector Space Model =====
#1 Cosine Similarity: 1.0000
-----
Document ID : 0
Title : Nucleosynthesis in magnetorotational supernovae: impact of the magnetic field configuration
Authors : M. Reichert, M. Bugli, J. Guillet, M. Obergaulinger, M. Á. Aloy, A. Arcones
Subjects : High Energy Astrophysical Phenomena
Abstract : The production of heavy elements is one of the main by-products of the explosive end of massive stars. A long sought goal is finding differentiated patterns in the nucleosynthesis yields, which could permit identifying a number of properties of the explosive core. Among them, the traces of the magnetic field topology are particularly important for \emph{extreme} supernova explosions, most likely hosted by magnetorotational effects. We investigate the nucleosynthesis of five state-of-the-art magnetohydrodynamic models with fast rotation that have been previously calculated in full 3D and that involve an accurate neutrino transport (M1). One of the models does not contain any magnetic field and synthesizes elements around the iron group, in agreement with other CC-SNe models in literature. All other models host a strong magnetic field of the same intensity, but with different topology. For the first time, we investigate the nucleosynthesis of MR-SNe models with a quadrupolar magnetic field and a 90 degree tilted dipole. We obtain a large variety of ejecta compositions reaching from iron nuclei to nuclei up to the third r-process peak. We assess the robustness of our results by considering the impact of different nuclear physics uncertainties such as different nuclear masses, $\beta^{\{-\}}$-decays and $\beta^{\{+\}}$-delayed neutron emission probabilities, neutrino reactions, fission, and a feedback of nuclear energy on the temperature. We find that the qualitative results do not change with different nuclear physics input. The properties of the explosion dynamics and the magnetic field configuration are the dominant factors determining the ejecta composition.
Comments :
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14402
-----
#2 Cosine Similarity: 0.1291
-----
Document ID : 70
Title : Magnetic fields of protoplanetary disks
Authors : Sergey A. Khaibrakhmanov
Subjects : Solar and Stellar Astrophysics, Earth and Planetary Astrophysics, Plasma Physics
Abstract : We review the current status of studies on accretion and protoplanetary disks of young stars with large-scale
=====
Console 1/A ×
electric and magnetic charges has been proved. Using conformal positive energy theorem, as well as, the positive mass theorem and adequate conformal transformations, we envisage the two alternative ways of proving that the exterior region of a certain radius of the studied static \{it photon sphere\}, is characterized by ADM mass, electric and magnetic charges.
Comments : 22 pages, RevTeX, to be published in Phys.Rev.D15
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14116
-----
#19 Cosine Similarity: 0.0453
-----
Document ID : 133
Title : A distribution-guided Mapper algorithm
Authors : Yuyang Tao, Shufei Ge
Subjects : Algebraic Topology, Machine Learning, Quantitative Methods
Abstract : Motivation: The Mapper algorithm is an essential tool to explore shape of data in topology data analysis. With a dataset as an input, the Mapper algorithm outputs a graph representing the topological features of the whole dataset. This graph is often regarded as an approximation of a reeb graph of data. The classic Mapper algorithm uses fixed interval lengths and overlapping ratios, which might fail to reveal subtle features of data, especially when the underlying structure is complex.
Results: In this work, we introduce a distribution guided Mapper algorithm named D-Mapper, that utilizes the property of the probability model and data intrinsic characteristics to generate density guided covers and provides enhanced topological features. Our proposed algorithm is a probabilistic model-based approach, which could serve as an alternative to non-probabilistic ones. Moreover, we introduce a metric accounting for both the quality of overlap clustering and extended persistence homology to measure the performance of Mapper type algorithm. Our numerical experiments indicate that the D-Mapper outperforms the classical Mapper algorithm in various scenarios. We also apply the D-Mapper to a SARS-COV-2 coronavirus RNA sequences dataset to explore the topological structure of different virus variants. The results indicate that the D-Mapper algorithm can reveal both vertical and horizontal evolution processes of the viruses.
Availability: Our package is available at https://github.com/ShufeiGe/D-Mapper.
Comments :
Date : 19 January 2024
PDF_URL : https://arxiv.org/pdf/2401.12237
-----
#20 Cosine Similarity: 0.0429
-----
Document ID : 138
Title : Approximating a linear dynamical system from non-sequential data
Authors : Cliff Stein, Pratik Worah
Subjects : Genomics
Abstract : Given non-sequential snapshots from instances of a dynamical system, we design a compressed sensing based algorithm that reconstructs the dynamical system. We formally prove that successful reconstruction is possible under the assumption that we can construct an approximate clock from a subset of the coordinates of the underlying system.
As an application, we argue that our assumption is likely true for genomic datasets, and we recover the underlying nuclear receptor networks and predict pathways, as opposed to genes, that may differentiate phenotypes in some publicly available datasets.
Comments :
Date : 22 January 2024
PDF_URL : https://arxiv.org/pdf/2401.11858
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

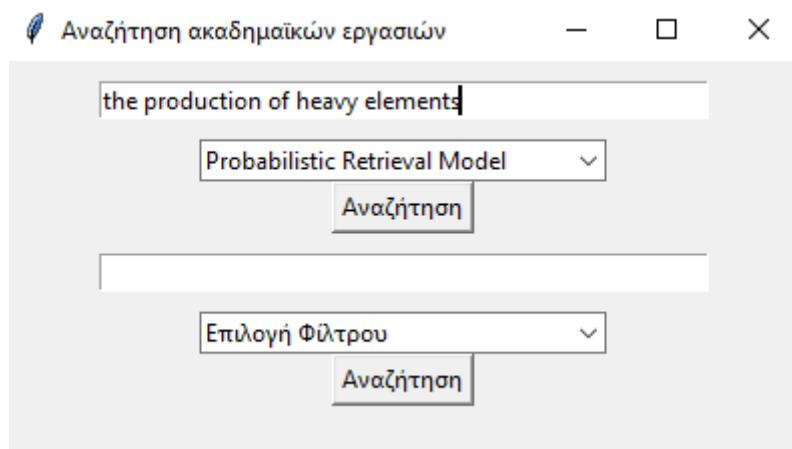
Εικόνα 4.B.6 Ταυτόσημο ερώτημα χρήστη με κάποιο κείμενο (VSM)

Probabilistic Retrieval Model

Η υλοποίηση του Probabilistic Retrieval Model πραγματοποιήθηκε στο ίδιο module (search_engine.py) με την [διεπαφή χρήστη](#) και συγκεκριμένα στην ίδια κλάση. Σαν είσοδο, ο αλγόριθμος παίρνει το ερώτημα χρήστη και το ανεστραμμένο ευρετήριο και οι υλοποιήσεις που λαμβάνουν χώρα είναι οι εξής:

- Προεπεξεργασία κειμένου: [Βίμα 2](#)
- Υπολογισμός TF-IDF των όρων των κειμένων: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Υπολογισμός TF-IDF των όρων του ερωτήματος: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Υπολογισμός του σκορ Okapi BM25 μεταξύ των κειμένων και του ερωτήματος: [Κατάταξη αποτελεσμάτων \(Ranking\)](#)
- Κατάταξη αποτελεσμάτων (Ranking)

Μερικές ενδεικτικές μελέτες περιπτώσεων από ερωτήματα χρηστών είναι οι εξής:



ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
Console 1/A X
=====
Ερώτημα ανάζησης : the production of heavy elements =====
Αλγόριθμος ανάκτησης : Probabilistic Retrieval Model =====

#1 BM25 Score: 41.4798
-----
Document ID : 211
Title : Optimal Queueing Regimes
Authors : Marco Scarsini, Eran Shmaya
Subjects : Theoretical Economics, Computer Science and Game Theory, Probability
Abstract : We consider an M/M/1 queueing model where customers can strategically decide whether to join the queue or balk and when to renege. We characterize the class of queueing regimes such that, for any parameters of the model, the socially efficient behavior is an equilibrium outcome.
Comments : MSC Class: 91A40; 60J28
Date : 24 January 2024
PDF_URL : https://arxiv.org/pdf/2401.13812
-----
#2 BM25 Score: 41.0244
-----
Document ID : 194
Title : The Interplay Between Logical Phenomena and the Cognitive System of the Mind
Authors : Kazem Haghnejad Azar
Subjects : Neurons and Cognition
Abstract : In this article, we employ mathematical concepts as a tool to examine the phenomenon of consciousness experience and logical phenomena. Through our investigation, we aim to demonstrate that our experiences, while not confined to limitations, cannot be neatly encapsulated within a singular collection. Our conscious experience emerges as a result of the developmental and augmentative trajectory of our cognitive system. As our cognitive abilities undergo refinement and advancement, our capacity for logical thinking likewise evolves, thereby manifesting a heightened level of conscious experience. The primary objective of this article is to embark upon a profound exploration of the concept of logical experience, delving into the intricate process by which these experiences are derived from our mind.
Comments :
Date : 21 January 2024
PDF_URL : https://arxiv.org/pdf/2401.09465
-----
#3 BM25 Score: 40.8304
-----
Document ID : 256
Title : The outcomes of generative AI are exactly the Nash equilibria of a non-potential game
Authors : Boualem Djehiche, Hamidou Tembine
Subjects : Computer Science and Game Theory
Abstract : In this article we show that the asymptotic outcomes of both shallow and deep neural networks such as those used in BloombergGPT to generate economic time series are exactly the Nash equilibria of a non-potential game. We then design and analyze deep neural network algorithms that converge to these equilibria. The methodology is extended to federated deep neural networks between clusters of regional servers and on-device clients. Finally, the variational inequalities behind large language models including encoder-decoder related transformers are established.
Comments : 24 pages. Accepted and to appear in: International Econometric Conference of Vietnam
Date : 22 January 2024
PDF_URL : https://arxiv.org/pdf/2401.12321
-----
Console 1/A X
PDF_URL : https://arxiv.org/pdf/2401.14000
-----
#19 BM25 Score: 39.5568
-----
Document ID : 189
Title : Dimensional Neuroimaging Endophenotypes: Neurobiological Representations of Disease Heterogeneity Through Machine Learning
Authors : Junhao Wen, Mathilde Antoniades, Zhijian Yang, Gyujoon Hwang, Ioanna Skampardonis, Rongguang Wang, Christos Davatzikos
Subjects : Machine Learning, Image and Video Processing, Quantitative Methods
Abstract : Machine learning has been increasingly used to obtain individualized neuroimaging signatures for disease diagnosis, prognosis, and response to treatment in neuropsychiatric and neurodegenerative disorders. Therefore, it has contributed to a better understanding of disease heterogeneity by identifying disease subtypes that present significant differences in various brain phenotypic measures. In this review, we first present a systematic literature overview of studies using machine learning and multimodal MRI to unravel disease heterogeneity in various neuropsychiatric and neurodegenerative disorders, including Alzheimer disease, schizophrenia, major depressive disorder, autism spectrum disorder, multiple sclerosis, as well as their potential in transdiagnostic settings. Subsequently, we summarize relevant machine learning methodologies and discuss an emerging paradigm which we call dimensional neuroimaging endophenotype (DNE). DNE dissects the neurobiological heterogeneity of neuropsychiatric and neurodegenerative disorders into a low dimensional yet informative, quantitative brain phenotypic representation, serving as a robust intermediate phenotype (i.e., endophenotype) largely reflecting underlying genetics and etiology. Finally, we discuss the potential clinical implications of the current findings and envision future research avenues.
Comments :
Date : 17 January 2024
PDF_URL : https://arxiv.org/pdf/2401.09517
-----
#20 BM25 Score: 39.5309
-----
Document ID : 19
Title : Uncovering Heterogeneity of Solar Flare Mechanism With Mixture Models
Authors : Bach Viet Do, Yang Chen, XuanLong Nguyen, Ward Manchester
Subjects : Solar and Stellar Astrophysics, Applications, Methodology
Abstract : The physics of solar flares occurring on the Sun is highly complex and far from fully understood. However, observations show that solar eruptions are associated with the intense kilogauss fields of active regions, where free energies are stored with field-aligned electric currents. With the advent of high-quality data sources such as the Geostationary Operational Environmental Satellites (GOES) and Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI), recent works on solar flare forecasting have been focusing on data-driven methods. In particular, black box machine learning and deep learning models are increasingly adopted in which underlying data structures are not modeled explicitly. If the active regions indeed follow the same laws of physics, there should be similar patterns shared among them, reflected by the observations. Yet, these black box models currently used in the literature do not explicitly characterize the heterogeneous nature of the solar flare data, within and between active regions. In this paper, we propose two finite mixture models designed to capture the heterogeneous patterns of active regions and their associated solar flare events. With extensive numerical studies, we demonstrate the usefulness of our proposed method for both resolving the sample imbalance issue and modeling the heterogeneity for rare energetic solar flare events.
Comments :
Date : 25 January 2024
PDF_URL : https://arxiv.org/pdf/2401.14345
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Εικόνα 4.B.7 Μικρό ερώτημα χρήστη (PRM)

Αναζήτηση ακαδημαϊκών εργασιών

s by which these experiences are derived from our mind

Probabilistic Retrieval Model

Αναζήτηση

Επιλογή Φίλτρου

Αναζήτηση

Console 1/A

```
===== Ερώτημα αναζήτησης : In this article, we employ mathematical concepts as a tool to examine the phenomenon of consciousness experience and logical phenomena. Through our investigation, we aim to demonstrate that our experiences, while not confined to limitations, cannot be neatly encapsulated within a singular collection. Our conscious experience emerges as a result of the developmental and augmentative trajectory of our cognitive system. As our cognitive abilities undergo refinement and advancement, our capacity for logical thinking likewise evolves, thereby manifesting a heightened level of conscious experience. The primary objective of this article is to embark upon a profound exploration of the concept of logical experience, delving into the intricate process by which these experiences are derived from our mind. =====
===== Αλγόριθμος ανάκτησης : Probabilistic Retrieval Model =====
#1 BM25 Score: 719.5806
-----
Document ID : 194
Title      : The Interplay Between Logical Phenomena and the Cognitive System of the Mind
Authors    : Kazem Haghnejad Azar
Subjects   : Neurons and Cognition
Abstract   : In this article, we employ mathematical concepts as a tool to examine the phenomenon of consciousness experience and logical phenomena. Through our investigation, we aim to demonstrate that our experiences, while not confined to limitations, cannot be neatly encapsulated within a singular collection. Our conscious experience emerges as a result of the developmental and augmentative trajectory of our cognitive system. As our cognitive abilities undergo refinement and advancement, our capacity for logical thinking likewise evolves, thereby manifesting a heightened level of conscious experience. The primary objective of this article is to embark upon a profound exploration of the concept of logical experience, delving into the intricate process by which these experiences are derived from our mind.
Comments   :
Date       : 21 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.09465
-----
#2 BM25 Score: 711.1157
-----
Document ID : 144
Title      : Accelerating Seed Location Filtering in DNA Read Mapping Using a Commercial Compute-in-SRAM Architecture
Authors    : Courtney Golden, Dan Ilan, Nicholas Cebry, Christopher Batten
Subjects   : Hardware Architecture, Genomics
Abstract   : DNA sequence alignment is an important workload in computational genomics. Reference-guided DNA assembly involves aligning many read sequences against candidate locations in a long reference genome. To reduce the computational load of this alignment, candidate locations can be pre-filtered using simpler alignment algorithms like edit distance. Prior work has explored accelerating filtering on simulated compute-in-DRAM, due to the massive parallelism of compute-in-memory architectures. In this paper, we present work-in-progress on accelerating filtering using a commercial compute-in-SRAM accelerator. We leverage the recently released Gemini accelerator platform from GSI Technology, which is the first, to our knowledge, commercial-scale compute-in-SRAM system. We accelerate the Myers' bit-parallel edit distance algorithm, producing average speedups of 14.1x over single-core CPU performance. Individual query/candidate alignments produce speedups of up to 24.1x. These early results suggest this novel architecture is well-suited to accelerating the filtering step of sequence-to-sequence DNA alignment.
Comments   : Journal ref: 5th Workshop on Accelerator Architecture in Computational Biology and Bioinformatics (AACBB), June 2023
Date       : 21 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.11685
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
Console 1/A ×

#18 BM25 Score: 687.6884
-----
Document ID : 10
Title       : Efficient Optimisation of Physical Reservoir Computers using only a Delayed Input
Authors     : Enrico Picco, Lina Jaurigue, Kathy Lüdge, Serge Massar
Subjects    : Emerging Technologies, Artificial Intelligence, Neural and Evolutionary Computing, Optics
Abstract    : We present an experimental validation of a recently proposed optimization technique for reservoir computing, using an optoelectronic setup. Reservoir computing is a robust framework for signal processing applications, and the development of efficient optimization approaches remains a key challenge. The technique we address leverages solely a delayed version of the input signal to identify the optimal operational region of the reservoir, simplifying the traditionally time-consuming task of hyperparameter tuning. We verify the effectiveness of this approach on different benchmark tasks and reservoir operating conditions.
Comments   :
Date       : 25 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14371

#19 BM25 Score: 687.0624
-----
Document ID : 251
Title       : Moen Meets Rotemberg: An Earthly Model of the Divine Coincidence
Authors     : Pascal Michaillat, Emmanuel Saez
Subjects    : Theoretical Economics
Abstract    : This paper proposes a model of the divine coincidence, explaining its recent appearance in US data. The divine coincidence matters because it helps explain the behavior of inflation after the pandemic, and it guarantees that the full-employment and price-stability mandates of the Federal Reserve coincide. In the model, a Phillips curve relating unemployment to inflation arises from Moen's (1997) directed search. The Phillips curve is nonvertical thanks to Rotemberg's (1982) price-adjustment costs. The model's Phillips curve guarantees that the rate of inflation is on target whenever the rate of unemployment is efficient, generating the divine coincidence. If we assume that wage decreases -- which reduce workers' morale -- are more costly to producers than price increases -- which upset customers -- the Phillips curve also displays a kink at the point of divine coincidence.
Comments   :
Date       : 22 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.12475

#20 BM25 Score: 685.8325
-----
Document ID : 211
Title       : Optimal Queueing Regimes
Authors     : Marco Scarsini, Eran Shmaya
Subjects    : Theoretical Economics, Computer Science and Game Theory, Probability
Abstract    : We consider an M/M/1 queueing model where customers can strategically decide whether to join the queue or balk and when to renege. We characterize the class of queueing regimes such that, for any parameters of the model, the socially efficient behavior is an equilibrium outcome.
Comments   : MSC Class: 91A40; 60J28
Date       : 24 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.13812
```

Εικόνα 4.B.8 Ταυτόσημο ερώτημα χρήστη με κάποιο κείμενο (PRM)

4.β.3 Αξιολόγηση

Boolean Retrieval

Ο κώδικας του Boolean Retrieval είναι υλοποιημένος με τέτοιο τρόπο, ώστε να μπορεί να αναζητήσει τόσο με απλές όσο και με σύνθετες εκφράσεις που θα του δοθούν από τον χρήστη μέσω ενός query αναζήτησης. Ωστόσο, οι συγκεκριμένες λέξεις-κλειδιά που είναι απαραίτητες για την λειτουργία του αλγορίθμου (and, or, not) πρέπει να είναι γραμμένες σωστά όσον αφορά την ορθογραφία τους, διότι το αν είναι με πεζά ή κεφαλαία τα γράμματα που θα εισάγει ο χρήστης δεν επηρεάζει σε κάτι την αναζήτηση. Στην περίπτωση που κάποια λέξη που δοθεί μέσω query από τον χρήστη δεν υπάρχει στο ευρετήριο, ο κώδικας θα μας χειρίστει το λάθος όπως και σε άλλα search engines αν μία λέξη δεν είναι εντός των ευρετηρίων, δεν υπάρχουν και τα σχετικά αποτελέσματα.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Vector Space Model

Ο κώδικας του Vector Space Model αποδίδει για το μικρό αποθετήριο της μηχανής αναζήτησης και αποτελεί σημαντικό εργαλείο για την ανάκτηση των εργασιών. Ο υπολογισμός συνημιτόνων αποτελεί προηγμένη τεχνική κατάταξης, οπότε, τα αποτελέσματα της ανάκτησης θα μπορούσαν να περιοριστούν στον απλό αλγόριθμο κατάταξης TF-IDF. Αξίζει να σημειωθεί, ότι δεν χρησιμοποιήθηκε η βιβλιοθήκη της Python, scikit-learn, η οποία παρέχει έτοιμες μεθόδους που δημιουργούν τα διανύσματα και υπολογίζουν τα συνημίτονα.

Probabilistic Retrieval Model

Ο κώδικας του Probabilistic Retrieval Model αποδίδει και για μεγάλα αποθετήρια της μηχανής αναζήτησης και αποτελεί σημαντικό εργαλείο για την ανάκτηση των εργασιών. Ο υπολογισμός του bm25 score αποτελεί προηγμένη τεχνική κατάταξης, όπου προσφέρει αποτελεσματικότητα στην ανάκτηση των εργασιών, αφαιρώντας με τους σταθερούς συντελεστές κορεσμού b και k, τους όρους που εμφανίζονται πολύ συχνά στην συλλογή και συνήθως αποτελούν μικρής σημασίας πληροφορία.

4.γ. Φιλτράρισμα αποτελεσμάτων αναζήτησης με διάφορα κριτήρια

4.γ.1 Σχεδιασμός

Το φιλτράρισμα των αποτελεσμάτων είναι μία επιλογή που έχει ο χρήστης μετά την εισαγωγή του query που θέλει να αναζητήσει και του αλγόριθμου που επιλέγει για να κάνει την αναζήτηση. Ουσιαστικά προσφέρει στον χρήστη την επιλογή από τα αποτελέσματα που πήρε να κάνει ένα φιλτράρισμα σε αυτά είτε βάση των/την συγγραφέα τους, είτε βάση της ημερομηνίας που δημοσιεύτηκαν. Αυτό επιτυγχάνεται μέσω ενός δεύτερου query στο οποίο ο χρήστης πρέπει να εισάγει όνομα συγγραφέα (στην περίπτωση που έχει επιλεγεί ως φιλτράρισμα το πεδίο “Συγγραφείς”) ώστε να του εμφανιστούν από τα αποτελέσματα που πήρε, μόνο αυτά στα οποία στους συγγραφέis εμπεριέχεται και αυτός/αυτή που ζητά. Αντίστοιχα το ίδιο πρέπει να κάνει και στην περίπτωση που θέλει να γίνει το φιλτράρισμα κατά ημερομηνία, ωστόσο πρέπει το πεδίο να είναι στο φιλτράρισμα κατά “Ημερομηνία”.

4.γ.2 Υλοποίηση

Για την υλοποίηση του φιλτραρίσματος χρειάστηκε πρόσβαση σε όλα τα δεδομένα που έχουμε σε json μορφή καθώς επίσης και ο αλγόριθμος που χρησιμοποιήθηκε για την αναζήτηση. Γνωρίζοντας τον αλγόριθμο που χρησιμοποιήθηκε η συνάρτηση filtering ξέρει από ποιόν πίνακα πρέπει να πάρει τα αποτελέσματα (boolean_results αν χρησιμοποιήθηκε boolean retrieval κλπ). Αυτά τα αποτελέσματα είναι τα doc_id των κειμένων που εκτυπώνονται με τον κάθε αλγόριθμο. Η συνάρτηση που ασχολείται με το φιλτράρισμα λαμβάνει το φίλτρο που επιλέχθηκε από τον χρήστη (ημερομηνία, συγγραφέας) και επίσης ένα query που περιέχει αυτό

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

που θέλει να ψάξει ο χρήστης στα αποτελέσματα. Αργότερα γίνεται ο απαραίτητος έλεγχος με τα πεδία author ή date και εκτυπώνονται τα κείμενα που έχουν ίδια πληροφορία με το query του χρήστη.

Μερικές ενδεικτικές μελέτες περιπτώσεων από ερωτήματα χρηστών είναι οι εξής:

The screenshot shows a search interface with two search bars. The top bar has the query "the product of heavy elements" and a dropdown menu set to "Vector Space Model". Below it is a search button labeled "Αναζήτηση". The bottom bar has the query "Jonathan D'Emidio" and a dropdown menu set to "Συγγραφείς". Below it is another search button labeled "Αναζήτηση".

Console 1/A

```
Φιλτράρισμα αποτελεσμάτων κατά: Συγγραφείς
-----
#1 Cosine Similarity: 0.0000
-----
Document ID : 2
Title      : Entanglement entropy and deconfined criticality: emergent SO(5) symmetry and proper lattice bipartition
Authors    : Jonathan D'Emidio, Anders W. Sandvik
Subjects   : Strongly Correlated Electrons, High Energy Physics - Lattice
Abstract   : We study the Rényi entanglement entropy (EE) of the two-dimensional  $J=Q$  model, the emblematic quantum spin model of deconfined criticality at the phase transition between antiferromagnetic and valence-bond-solid ground states. Quantum Monte Carlo simulations with an improved EE scheme reveal critical corner contributions that scale logarithmically with the system size, with a coefficient in remarkable agreement with the form expected from a large- $N$  conformal field theory with  $SO(N=5)$  symmetry. However, details of the bipartition of the lattice are crucial in order to observe this behavior. If the subsystem for the reduced density matrix does not properly accommodate valence-bond fluctuations, logarithmic contributions appear even for corner-less bipartitions. We here use a 45° tilted cut on the square lattice. Beyond supporting an  $SO(5)$  deconfined quantum critical point, our results for both the regular and tilted cuts demonstrate important microscopic aspects of the EE that are not captured by conformal field theory.
Comments   : 5 pages, 3 figures
Date       : 25 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14396
```

Εικόνα 4.Γ.1 Φιλτράρισμα με κριτήριο των Συγγραφέα

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

🔍 Αναζήτηση ακαδημαϊκών εργασιών

the product of heavy elements

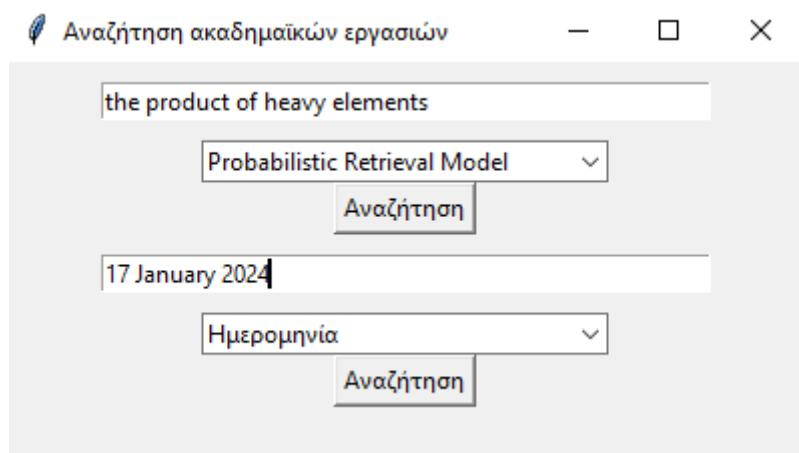
Probabilistic Retrieval Model

Αναζήτηση

17 January 2024

Ημερομηνία

Αναζήτηση



ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
Console 1/A ×
Ωιλτράρισμα αποτελεσμάτων κατά: Ημερομηνία
-----
#1 BM25 Score: 39.8124
-----
Document ID : 190
Title      : Is the Emergence of Life an Expected Phase Transition in the Evolving Universe?
Authors    : Stuart Kauffman, Andrea Roli
Subjects   : Populations and Evolution, Biological Physics
Abstract   : We propose a novel definition of life in terms of which its emergence in the universe is expected, and its ever-creative open-ended evolution is entailed by no law. Living organisms are Kantian Wholes that achieve Catalytic Closure, Constraint Closure, and Spatial Closure. We here unite for the first time two established mathematical theories, namely Collectively Autocatalytic Sets and the Theory of the Adjacent Possible. The former establishes that a first-order phase transition to molecular reproduction is expected in the chemical evolution of the universe where the diversity and complexity of molecules increases; the latter posits that, under loose hypotheses, if the system starts with a small number of beginning molecules, each of which can combine with copies of itself or other molecules to make new molecules, over time the number of kinds of molecules increases slowly but then explodes upward hyperbolically. Together these theories imply that life is expected as a phase transition in the evolving universe. The familiar distinction between software and hardware loses its meaning in living cells. We propose new ways to study the phylogeny of metabolisms, new astronomical ways to search for life on exoplanets, new experiments to seek the emergence of the most rudimentary life, and the hint of a coherent testable pathway to prokaryotes with template replication and coding.
Comments   :
Date       : 17 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.09514
-----
#2 BM25 Score: 39.5568
-----
Document ID : 189
Title      : Dimensional Neuroimaging Endophenotypes: Neurobiological Representations of Disease Heterogeneity Through Machine Learning
Authors    : Junhao Wen, Mathilde Antoniades, Zhijian Yang, Gyujoon Hwang, Ioanna Skampardonis, Rongguang Wang, Christos Davatzikos
Subjects   : Machine Learning, Image and Video Processing, Quantitative Methods
Abstract   : Machine learning has been increasingly used to obtain individualized neuroimaging signatures for disease diagnosis, prognosis, and response to treatment in neuropsychiatric and neurodegenerative disorders. Therefore, it has contributed to a better understanding of disease heterogeneity by identifying disease subtypes that present significant differences in various brain phenotypic measures. In this review, we first present a systematic literature overview of studies using machine learning and multimodal MRI to unravel disease heterogeneity in various neuropsychiatric and neurodegenerative disorders, including Alzheimer disease, schizophrenia, major depressive disorder, autism spectrum disorder, multiple sclerosis, as well as their potential in transdiagnostic settings. Subsequently, we summarize relevant machine learning methodologies and discuss an emerging paradigm which we call dimensional neuroimaging endophenotype (DNE). DNE dissects the neurobiological heterogeneity of neuropsychiatric and neurodegenerative disorders into a low dimensional yet informative, quantitative brain phenotypic representation, serving as a robust intermediate phenotype (i.e., endophenotype) largely reflecting underlying genetics and etiology. Finally, we discuss the potential clinical implications of the current findings and envision future research avenues.
Comments   :
Date       : 17 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.09517
-----
Console 1/A ×
PDF_URL   : https://arxiv.org/pdf/2401.09255
-----
#16 BM25 Score: 35.4404
-----
Document ID : 295
Title      : Early Prediction of Geomagnetic Storms by Machine Learning Algorithms
Authors    : Iris Yan
Subjects   : Machine Learning
Abstract   : Geomagnetic storms (GS) occur when solar winds disrupt Earth's magnetosphere. GS can cause severe damages to satellites, power grids, and communication infrastructures. Estimate of direct economic impacts of a large scale GS exceeds $40 billion a day in the US. Early prediction is critical in preventing and minimizing the hazards. However, current methods either predict several hours ahead but fail to identify all types of GS, or make predictions within short time, e.g., one hour ahead of the occurrence. This work aims to predict all types of geomagnetic storms reliably and as early as possible using big data and machine learning algorithms. By fusing big data collected from multiple ground stations in the world on different aspects of solar measurements and using Random Forests regression with feature selection and downsampling on minor geomagnetic storm instances (which carry majority of the data), we are able to achieve an accuracy of 82.55% on data collected in 2021 when making early predictions three hours in advance. Given that important predictive features such as historic Kp indices are measured every 3 hours and their importance decay quickly with the amount of time in advance, an early prediction of 3 hours ahead of time is believed to be close to the practical limit.
Comments   : 14 pages, 7 figures
Date       : 17 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.10290
-----
#17 BM25 Score: 34.9005
-----
Document ID : 187
Title      : Functional Linear Non-Gaussian Acyclic Model for Causal Discovery
Authors    : Tian-Le Yang, Kuang-Yao Lee, Kun Zhang, Joe Suzuki
Subjects   : Machine Learning, Statistics Theory, Neurons and Cognition, Methodology
Abstract   : In causal discovery, non-Gaussianity has been used to characterize the complete configuration of a Linear Non-Gaussian Acyclic Model (LiNGAM), encompassing both the causal ordering of variables and their respective connection strengths. However, LiNGAM can only deal with the finite-dimensional case. To expand this concept, we extend the notion of variables to encompass vectors and even functions, leading to the Functional Linear Non-Gaussian Acyclic Model (Func-LiNGAM). Our motivation stems from the desire to identify causal relationships in brain-effective connectivity tasks involving, for example, fMRI and EEG datasets. We demonstrate why the original LiNGAM fails to handle these inherently infinite-dimensional datasets and explain the availability of functional data analysis from both empirical and theoretical perspectives. {We establish theoretical guarantees of the identifiability of the causal relationship among non-Gaussian random vectors and even random functions in infinite-dimensional Hilbert spaces.} To address the issue of sparsity in discrete time points within intrinsic infinite-dimensional functional data, we propose optimizing the coordinates of the vectors using functional principal component analysis. Experimental results on synthetic data verify the ability of the proposed framework to identify causal relationships among multivariate functions using the observed samples. For real data, we focus on analyzing the brain connectivity patterns derived from fMRI data.
Comments   :
Date       : 17 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.09641
```

Εικόνα 4.Γ.2 Φιλτράρισμα με κριτήριο την Ημερομηνία Δημοσίευσης

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

4.γ.3 Αξιολόγηση

Η συνάρτηση φιλτραρίσματος των αποτελεσμάτων λειτουργεί ορθά, ωστόσο πρόκειται για μία αρκετά απλοϊκή συνάρτηση η οποία απαιτεί από τον χρήστη να εισάγει με ακρίβεια τα στοιχεία που ενδιαφέρεται να ψάξει στα αποτελέσματα των αλγορίθμων. Η μόνη περίπτωση κατά την οποία η συνάρτηση δεν λειτουργεί είναι όταν ως είσοδο στο φιλτράρισμα ο χρήστης εισάγει πάνω από 1 συγγραφέα. Ωστόσο, για ένα συγγραφέα η συνάρτηση πληροί τη λογική του φιλτρου.

Επεξεργασία ερωτήματος (Query Processing)

Σχεδιασμός

Η διαδικασία της επεξεργασίας ερωτημάτων (Query Processing) που ακολουθεί ο αλγόριθμος ανάκτησης [Boolean Retrieval](#), βασίζεται στην άλγεβρα Boole και τη θεωρία συνόλων, χρησιμοποιώντας τρεις βασικές Boolean λειτουργίες: OR, AND και NOT. Αυτές οι λειτουργίες επηρεάζουν τα αποτελέσματα της αναζήτησης ανάλογα με τη θέση τους στο ερώτημα (query) που λαμβάνεται από τον χρήστη. Η ανάκτηση των σχετικών εγγράφων γίνεται με την χρήση του ανεστραμμένου ευρετηρίου. Το θεωρητικό/μαθηματικό υπόβαθρο του σχεδιαστικού μοντέλου είναι το εξής:

AND

$$AND_RES = Q_0 \cap Q_1 = \{doc_{id} \mid doc_{id} \in Q_0 \text{ and } doc_{id} \in Q_1\}$$

AND_RES □ Τα κείμενα που περιέχουν και τους δύο όρους Q_0, Q_1

Q_0 □ Τα κείμενα που περιέχουν τον όρο Q_0

Q_1 □ Τα κείμενα που περιέχουν τον όρο Q_1

OR

$$OR_RES = Q_0 \cup Q_1 = \{doc_{id} \mid doc_{id} \in Q_0 \text{ or } doc_{id} \in Q_1\}$$

RES □ Τα κείμενα που περιέχουν είτε τον όρο Q_0 , είτε τον όρο Q_1 είτε και τους

Q_0 □ Τα κείμενα που περιέχουν τον όρο Q_0

Q_1 □ Τα κείμενα που περιέχουν τον όρο Q_1

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

NOT

$$NOT_RES = \neg Q_0 = \{doc_{id} \mid doc_{id} \in D \setminus Q_0\}$$

$RES \sqsubseteq$ Τα κείμενα που δεν περιέχουν τον όρο Q_0

$D \sqsubseteq$ Τα κείμενα της συλλογής

$Q_0 \sqsubseteq$ Τα κείμενα που περιέχουν τον όρο Q_0

Η διαδικασία πραγματοποιείται για απλά Boolean ερωτήματα χρήστη π.χ. black OR white, NOT black, black AND WHITE. Για σύνθετα ερωτήματα, ο αλγόριθμος ανάκτησης [Boolean Retrieval](#) καθορίζει την σειρά προτεραιότητας των πράξεων, όπου σπάει το σύνθετο ερώτημα σε απλά υπο-ερωτήματα και τα στέλνει ανάλογα με την σειρά προτεραιότητας, για επεξεργασία.

Υλοποίηση

Η υλοποίηση της επεξεργασίας ερωτήματος πραγματοποιήθηκε σε ξεχωριστό module (query_processing.py). Αρχικά, οι όροι του Boolean ερωτήματος διαχωρίζονται σε λεκτικές μονάδες και με χρήση του ανεστραμμένου ευρετηρίου από μία ρουτίνα, αντικαθίστονται όλες οι λέξεις-όροι (terms) του Boolean ερωτήματος με τις λίστες (posting lists) με τα κείμενα (doc_id) στα οποία εμφανίζονται. Οι Boolean πράξεις πραγματοποιούνται στις λίστες αυτές. Αν δεν υπάρχει ο όρος στο ευρετήριο, τότε η ρουτίνα χειρίζεται το λάθος.

Η ανάκτηση των σχετικών εγγράφων για κάθε λογικό τελεστή υλοποιείται ως εξής:

AND

Εφόσον ο τρέχον όρος είναι ο τελεστής «AND», ανακτώνται τα doc_id του αποθετηρίου που ανήκουν στην λίστα με τα doc_id που εμφανίζεται ο προηγούμενος και στην αντίστοιχη λίστα του επόμενου όρου.

OR

Εφόσον ο τρέχον όρος είναι ο τελεστής «OR», ανακτώνται τα doc_id του αποθετηρίου που ανήκουν στην λίστα με τα doc_id που εμφανίζεται ο προηγούμενος και τα doc_id που ανήκουν στην λίστα του επόμενου όρου.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

NOT

Εφόσον ο τρέχων όρος είναι ο τελεστής «NOT», ανακτώνται τα doc_id του αποθετηρίου που δεν ανήκουν στην λίστα με τα doc_id που εμφανίζεται ο επόμενος όρος

Η υλοποίηση επικεντρώνεται στη διαχείριση των Boolean λειτουργικών τελεστών AND, OR, NOT για την επεξεργασία ερωτημάτων, χρησιμοποιώντας σύνολα για τις πράξεις συνόλων.

Αξιολόγηση

Η επεξεργασία ερωτήματος αποτελεί ένα χρήσιμο εργαλείο για τις μηχανές αναζήτησης, καθώς, χειρίζεται τα ερωτήματα χρήστη που περιέχουν Boolean λειτουργίες (AND, OR, NOT) με βάση την άλγεβρα Boole και την θεωρία συνόλων που ακολουθεί ο αλγόριθμος ανάκτησης [Boolean Retrieval](#). Η φιλοσοφία «ανήκει στο x κείμενο, δεν ανήκει στο y κείμενο» παρέχει έναν αποτελεσματικό τρόπο ανάκτησης κειμένων βάσει του ερωτήματος που θέτει ο χρήστης. Ο [σχεδιασμός](#) παρουσιάζει το θεωρητικό υπόβαθρο με βάση την θεωρία των συνόλων, ενώ η [υλοποίηση](#) το πρακτικό υπόβαθρο με βάση την άλγεβρα Boole. Η αντικατάσταση των όρων-λέξεων με τις λίστες με τα κείμενα της συλλογής που ανήκουν, σύμφωνα με το ανεστραμμένο ευρετήριο, επιταχύνει την διαδικασία της ανάκτησης των κειμένων που θέτει σαν ερώτημα ο χρήστης. Η αναπαράσταση της διαδικασίας με την θεωρία των συνόλων και την άλγεβρα Boole καθιστούν τις ανακτήσεις των κειμένων αποτελεσματικές.

Κατάταξη αποτελεσμάτων (Ranking)

Σχεδιασμός

Το σχεδιαστικό μοντέλο της κατάταξης αποτελεσμάτων (Ranking) ακολουθεί αρχικά την εφαρμογή του απλού αλγορίθμου κατάταξης TF-IDF (Term Frequency-Inverse Document Frequency) και ύστερα την εφαρμογή των προηγμένων τεχνικών κατάταξης Cosine Similarity που ακολουθεί ο αλγόριθμος ανάκτησης [Vector Space Model](#) και Okapi BM25 που ακολουθεί ο αλγόριθμος ανάκτησης [Probabilistic Retrieval Model](#). Οι προαναφερόμενες εφαρμογές συμβάλλουν στην εκτίμηση της σημασιολογικής σημασίας των όρων σ' ένα σύνολο από κείμενα, όπου με βάση ενός ερωτήματος χρήστη, υπολογίζονται οι συντελεστές ομοιότητας συνημιτόνων (cosine similarity) ή οι συντελεστές BM25 (BM25 Score) μεταξύ των όρων του ερωτήματος χρήστη και των όρων των κειμένων της συλλογής. Το μοντέλο του TF-IDF ακολουθεί τον υπολογισμό του γινομένου $TF \times IDF$ για κάθε όρο της συλλογής και για κάθε όρο του ερωτήματος χρήστη, ενώ τα μοντέλα Cosine Similarity και Okapi BM25 ακολουθούν τον υπολογισμό του συντελεστή ομοιότητας συνημιτόνων και ενός σκορ αντίστοιχα με βάση αυτά τα δύο γινόμενα. Τέλος, τα αποτελέσματα αναζήτησης ταξινομούνται με

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

βάση των μεγαλύτερο συντελεστή. Το θεωρητικό/μαθηματικό υπόβαθρο του σχεδιαστικού μοντέλου είναι το εξής:

TF-IDF

$$TF_{t,d} = \frac{N_{t,d}}{\sum_k N_{k,d}}$$

$TF_{t,d}$ □ Η συχνότητα εμφάνισης του όρου t στο κείμενο d

$N_{t,d}$ □ Το πλήθος εμφανίσεων του όρου t στο κείμενο d

$\sum_k N_{k,d}$ □ Το πλήθος των όρων του κειμένου d

$$DF_t = |D_t|$$

DF_t □ Η συχνότητα εμφάνισης των κειμένων όπου εμφανίζεται ο όρος t

$|D_t|$ □ Ο αριθμός των κειμένων όπου εμφανίζεται ο όρος t

$$IDF_t = \log \log \frac{|D|}{DF_t}$$

IDF_t □ Η αντίστροφη συχνότητα εμφάνισης των κειμένων όπου εμφανίζεται ο όρος t. Οι όροι με υψηλή συχνότητα εμφάνισης τους σε κείμενα αποτελούν μικρής σημασίας πληροφορία και αυτό οφείλεται από το γεγονός ότι ο λογάριθμος τείνει προς το 0.

$|D|$ □ Το πλήθος της συλλογής εργασιών

DF_t □ Η συχνότητα εμφάνισης των κειμένων όπου εμφανίζεται ο όρος t

Cosine Similarity

$$\cos \cos (q, d) = \frac{q \times d}{\|q\| \times \|d\|} \rightarrow \cos \cos (q, d) = \frac{\sum_{t=1}^n (q_t \times d_t)}{\sqrt{\sum_{t=1}^n (q_t)^2} \times \sqrt{\sum_{t=1}^n (d_t)^2}}$$

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

$\sum_{t=1}^n (q_t \times d_t)$ □ Το εσωτερικό γινόμενο του διανύσματος του ερωτήματος χρήστη με το διάνυσμα κάθε

κειμένου της συλλογής

$\sqrt{\sum_{t=1}^n (q_t)^2}$ □ Η Ευκλείδια απόσταση του διανύσματος του ερωτήματος χρήστη

$\sqrt{\sum_{t=1}^n (d_t)^2}$ □ Η Ευκλείδια απόσταση του διανύσματος κάθε κειμένου της συλλογής

Okapi BM25

$$(q, d) = \sum_{t=1}^n IDF(q_t) \times \frac{TF(q_t, d) \times (k + 1)}{TF(q_t, d) + k \times (1 - b + b \times \frac{|d|}{avg_dlen})}$$

$\sum_{t=1}^n IDF(q_t)$ □ Η αντίστροφη συχνότητα εμφάνισης των κειμένων όπου εμφανίζονται οι όροι του ερωτήματος

$TF(q_t, d)$ □ Η συχνότητα εμφάνισης των όρων του ερωτήματος στο κείμενο d

k □ Θετική παράμετρος ($k = 1.2$) η οποία ελέγχει τους όρους που εμφανίζονται πολύ συχνά σ' ένα κείμενο (TF) και συνήθως, πρόκειται για μικρής σημασίας πληροφορία για το κείμενο αυτό

b □ Θετική παράμετρος ($b = 0.75$) η οποία ελέγχει τους όρους που εμφανίζονται πολύ συχνά σε κείμενα (DF) και συνήθως, πρόκειται για μικρής σημασίας πληροφορία για την συλλογή

$|d|$ □ Το μέγεθος του κειμένου d

avg_dlen □ Ο μέσος όρος μεγέθους της συλλογής

Υλοποίηση

Οι αλγόριθμοι κατάταξης υλοποιούνται σε ξεχωριστό module (ranking.py) και περιλαμβάνουν τις εξής ρουτίνες:

TF-IDF

- Υπολογισμός TF-IDF των όρων των εργασιών:
Για κάθε όρο της συλλογής εργασιών (απ' τα πεδία abstract):

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

1. Προεπεξεργασία κειμένου ([βήμα 2](#)) και διαχωρισμός σε λεκτικές μονάδες (tokenization)
 2. Υπολογισμός της συχνότητας εμφάνισης κάθε όρου στο κείμενο (Term Frequency)
 3. Υπολογισμός της συχνότητας εμφάνισης των κειμένων που εμφανίζεται ο κάθε όρος (Document Frequency)
 4. Υπολογισμός της αντίστροφης συχνότητας εμφάνισης των κειμένων που εμφανίζεται ο κάθε όρος (Inverse Document Frequency)
 5. Υπολογισμός της συχνότητας εμφάνισης κάθε όρου στο κείμενο (TF × IDF)
- Υπολογισμός TF-IDF των όρων του ερωτήματος χρήστη:
Για κάθε όρο του ερωτήματος χρήστη (query):
 1. Προεπεξεργασία του ερωτήματος (ρουτίνα του [βήματος 2](#)) και διαχωρισμός σε λεκτικές μονάδες (tokenization)
 2. Υπολογισμός της συχνότητας εμφάνισης κάθε όρου του ερωτήματος στο ερώτημα (Term Frequency)
 3. Υπολογισμός της αντίστροφης συχνότητας εμφάνισης των κειμένων που εμφανίζεται ο κάθε όρος του κειμένου (Inverse Document Frequency)
 4. Υπολογισμός της συχνότητας εμφάνισης κάθε όρου του ερωτήματος σε κάθε κείμενο (TF × IDF)

Cosine Similarity

- Υπολογισμός της ομοιότητας συνημιτόνων (Vector Space Model) μεταξύ του ερωτήματος χρήστη και των εργασιών (Cosine Similarity):
Για κάθε κείμενο της συλλογής, η ομοιότητα του με το ερώτημα χρήστη υπολογίζεται ως εξής:
 1. Υπολογισμός εσωτερικού γινομένου των διανυσμάτων του ερωτήματος χρήστη με το διάνυσμα κάθε κειμένου της συλλογής
 2. Υπολογισμός της Ευκλείδιας απόστασης του διανύσματος του ερωτήματος χρήστη
 3. Υπολογισμός της Ευκλείδιας απόστασης του διανύσματος κάθε κειμένου της συλλογής
 4. Υπολογισμός της ομοιότητας μεταξύ του ερωτήματος χρήστη και του κειμένου
- Κατάταξη των εργασιών με βάση την ομοιότητα συνημιτόνου

Okapi BM25

- Υπολογισμός του συντελεστή BM25 Score (Probabilistic retrieval model) μεταξύ του ερωτήματος χρήστη και των εργασιών:
Για κάθε κείμενο της συλλογής, η ομοιότητα του με το ερώτημα χρήστη υπολογίζεται ως εξής:
 1. Προεπεξεργασία του ερωτήματος (ρουτίνα του [βήματος 2](#))
 2. Υπολογισμός του μεγέθους της εργασίας
 3. Υπολογισμός μέσου όρου μεγέθους συλλογής εργασιών
 4. Υπολογισμός της συχνότητας εμφάνισης των κειμένων που εμφανίζεται ο κάθε όρος (Document Frequency)
 5. Υπολογισμός της αντίστροφης συχνότητας εμφάνισης των κειμένων που εμφανίζεται ο κάθε όρος (Inverse Document Frequency)
 6. Υπολογισμός συχνότητας εμφάνισης του όρου του ερωτήματος στο κάθε κείμενο (Term Frequency)
 7. Υπολογισμός BM25 συντελεστή για κάθε όρο του ερωτήματος

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Κατάταξη των εγγράφων με βάση τον συντελεστή BM25 Score

Αξιολόγηση

Οι αλγόριθμοι κατάταξης αποτελούν σημαντικά εργαλεία για τις μηχανές αναζήτησης ως προς την αποτελεσματικότητα της ανάκτησης εγγράφων. Ο απλός αλγόριθμος κατάταξης TF-IDF υπολογίζει το πόσο σημαντικός είναι ένας όρος στην συλλογή, ενώ οι προηγμένοι αλγόριθμοι Cosine Similarity και Okapi BM25 επικεντρώνονται στην ομοιότητα μεταξύ του ερωτήματος χρήστη και των κειμένων της συλλογής. Η φιλοσοφία «πόσο πλησιέστερα είναι κάποια κείμενα ως προς το ερώτημα χρήστη» παρέχει έναν αποτελεσματικό τρόπο ανάκτησης κειμένων βάσει του ερωτήματος που θέτει ο χρήστης.

5. Αξιολόγηση συστήματος

5.α. Σύνολα Δεδομένων (Dataset)

Η άντληση των δεδομένων γίνεται από το [arxiv.org](#) και με διάφορα επιστημονικά πεδία όπως είναι η Φυσική, τα Μαθηματικά κλπ. Ως ένας χώρος εναπόθεσης επιστημονικών εργασιών, παρέχει ένα μεγάλο αριθμό εγγράφων και λέξεων για το ευρετήριο.

5.β. Σενάρια Αξιολόγησης

Ο χρήστης μπορεί μέσω της μηχανής αναζήτησης να ψάξει να κάνει όποια αναζήτηση θέλει σε περιεχόμενα από ένα μεγάλο ευρετήριο, ώστόσο τα τελικά αποτελέσματα που θα εμφανίζονται είναι περιορισμένα σε μέγεθος μιας και να χωράνε στην κονσόλα του κάθε IDE.

5.γ. Βιβλιοθήκες Python

Οι βιβλιοθήκες που χρησιμοποιήθηκαν για την δημιουργία της μηχανής αναζήτησης είναι αρκετές, μερικές εξ' αυτών είναι η json που επιτρέπει να αποθηκευτούν δεδομένα σε συγκεκριμένη μορφή, η request η οποία επιτρέπει την ανάκτηση μιας σελίδας μέσω ενός HTTP-GET αιτήματος και η bs4 που επιστρέφει την σελίδα σε μη-δομημένη μορφή HTML (parsing)

5.δ. Εφαρμογές μετρικών

Η μηχανή αναζήτησης για να ξεκινήσει σε μία μέση δοκιμή της χρειάζεται περίπου και να αντλήσει τα κατάλληλα δεδομένα από το [arxiv.org](#) χρειάζεται κατά μέσο περίπου στα 12 δευτερόλεπτα. Τα paper τα οποία συλλέγει κυμαίνονται από 200 εώς και 800 γιατί όπως αναφέρθηκε στην [υλοποίηση](#) του web crawler, απευθύνεται σε 2-8 αντίστοιχους συνδέσμους για την ανάκτηση των δεδομένων που θα αποθηκευτούν στο αποθετήριο (dataset). Ενδεικτικά μέτρα αξιολόγησης είναι:

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Ακρίβεια (Precision) είναι το ποσοστό των ανακτηθέντων κειμένων που είναι συναφή

$$P = \frac{\# \text{συναφή ανακτηθέντα κείμενα}}{\# \text{ανακτηθέντα κείμενα}}$$

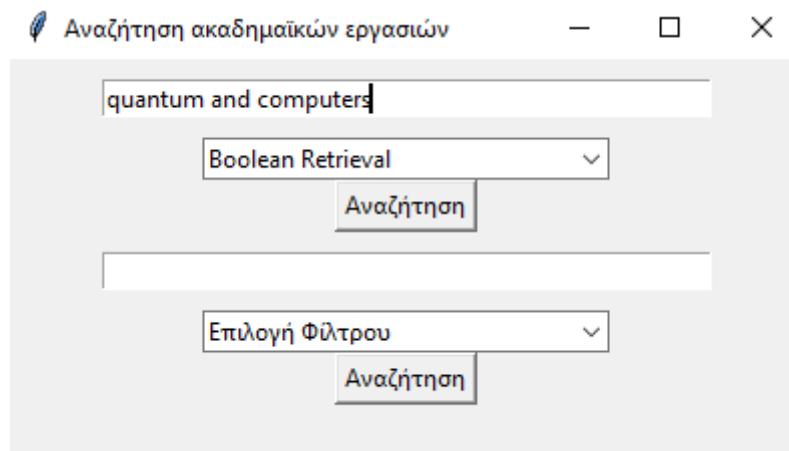
- Ανάκληση (Recall) είναι το ποσοστό των συναφών κειμένων που έχουν ανακτηθεί

$$R = \frac{\# \text{συναφή ανακτηθέντα κείμενα}}{\# \text{συναφή κείμενα}}$$

- F1 Score είναι το ποσοστό εξισορρόπησης της ακρίβειας και της ανάκλησης

$$\frac{1}{F} = \frac{1}{2} \times \left(\frac{1}{P} + \frac{1}{R} \right)$$

Ενδεικτική μελέτη περίπτωσης είναι η εξής:



ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
===== Ερώτημα αναζήτησης : quantum and computers =====
===== Αλγόριθμος ανάκτησης : Boolean Retrieval =====

#1
-----
Document ID : 4
Title      : Quantum types: going beyond qubits and quantum gates
Authors    : Tamás Varga, Yaiza Aragón-Soria, Manuel Oriol
Subjects   : Quantum Physics
Abstract   : Quantum computing is a growing field with significant potential applications. Learning how to code quantum programs means understanding how qubits work and learning to use quantum gates. This is analogous to creating classical algorithms using logic gates and bits. Even after learning all concepts, it is difficult to create new algorithms, which hinders the acceptance of quantum programming by most developers. This article outlines the need for higher-level abstractions and proposes some of them in a developer-friendly programming language called Rhyme. The new quantum types are extensions of classical types, including bits, integers, floats, characters, arrays, and strings. We show how to use such types with code snippets.
Comments   : 4 pages, accepted for the 5th International Workshop on Quantum Software Engineering (Q-SE 2024)
Date       : 26 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.15073
-----

#2
-----
Document ID : 7
Title      : Universality conditions of unified classical and quantum reservoir computing
Authors    : Francesco Monzani, Enrico Prati
Subjects   : Quantum Physics, Computational Physics
Abstract   : Reservoir computing is a versatile paradigm in computational neuroscience and machine learning, that exploits the non-linear dynamics of a dynamical system - the reservoir - to efficiently process time-dependent information. Since its introduction, it has exhibited remarkable capabilities in various applications. As widely known, classes of reservoir computers serve as universal approximators of functionals with fading memory. The construction of such universal classes often appears context-specific, but in fact, they follow the same principles. Here we present a unified theoretical framework and we propose a ready-made setting to secure universality. We test the result in the arising context of quantum reservoir computing. Guided by such a unified theorem we suggest why spatial multiplexing may serve as a computational resource when dealing with quantum registers, as empirically observed in specific implementations on quantum hardware. The analysis sheds light on a unified view of classical and quantum reservoir computing.
Comments   : 24 pages, 1 figure
Date       : 26 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.15067
-----

#3
-----
Document ID : 8
Title      : Efficient High-Dimensional Entangled State Analyzer with Linear Optics
Authors    : Niv Bharos, Liubov Markovich, Johannes Borregaard
Subjects   : Quantum Physics, Optics
Abstract   : The use of higher-dimensional photonic encodings (qudits) instead of two-dimensional encodings (qubits) can improve the loss tolerance and reduce the computational resources of photonic-based quantum information processing. To harness this potential, efficient schemes for entangling operations such as the high-dimensional generalization of a linear optics Bell measurement will be required. We show how an efficient high-dimensional entangled state analyzer can be implemented with linear optics and auxiliary photonic states. The Schmidt rank of the auxiliary state in our protocol scales only linearly with the dimensions of the input states instead of more than exponentially, as in previous proposals. In addition, we outline how the state can be generated deterministically from a single quantum emitter coupled to a small qubit processor. Our protocol thus outlines an experimentally feasible route for efficient, high-dimensional entangled state analyzers with linear optics.
Comments   : 10 pages, 5 figures
Date       : 26 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.15066
-----

#4
-----
Document ID : 42
Title      : Geometric measure of entanglement of quantum graph states prepared with controlled phase shift operators
Authors    : N. A. Suslovská
Subjects   : Quantum Physics
Abstract   : We consider graph states generated by the action of controlled phase shift operators on a separable state of a multi-qubit system. The case when all the qubits are initially prepared in arbitrary states is investigated. We obtain the geometric measure of entanglement of a qubit with the remaining system in graph states represented by arbitrary weighted graphs and establish its relationship with state parameters. For two-qubit graph states, the geometric measure of entanglement is also quantified on IBM's simulator Qiskit Aer and quantum processor ibmq lima based on auxiliary mean spin measurements. The results of quantum computations verify our analytical predictions.
Comments   : 17 pages, 7 Postscript figures
Date       : 26 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14997
-----

#5
-----
Document ID : 73
Title      : Shadow simulation of quantum processes
Authors    : Xuanqiang Zhao, Xin Wang, Giulio Chiribella
Subjects   : Quantum Physics
Abstract   : We introduce the task of shadow process simulation, where the goal is to reproduce the expectation values of arbitrary quantum observables at the output of a target physical process. When the sender and receiver share classical random bits, we show that the performance of shadow process simulation exceeds that of conventional process simulation protocols in a variety of scenarios including communication, noise simulation, and data compression. Remarkably, shadow simulation provides increased accuracy without any increase in the sampling cost. Overall, shadow simulation provides a unified framework for a variety of quantum protocols, including probabilistic error cancellation and circuit knitting in quantum computing.
Comments   : 21 pages, 4 figures
Date       : 26 January 2024
PDF_URL   : https://arxiv.org/pdf/2401.14934
```

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

```
-----  
#6  
-----  
Document ID : 84  
Title : Many-excitation removal of a transmon qubit using a single-junction quantum-circuit refrigerator and a two-tone microwave drive  
Authors : Wallace Teixeira, Timm Mörstedt, Arto Viitanen, Heidi Kivijärvi, András Gunyhó, Maaria Tiiri, Suman Kundu, Aashish Sah, Vasilii Vadimov, Mikko Möttönen  
Subjects : Quantum Physics  
Abstract : Achieving fast and precise initialization of qubits is a critical requirement for the successful operation of quantum computers. The combination of engineered environments with all-microwave techniques has recently emerged as a promising approach for the reset of superconducting quantum devices. In this work, we experimentally demonstrate the utilization of a single-junction quantum-circuit refrigerator (QCR) for an expeditious removal of several excitations from a transmon qubit. The QCR is indirectly coupled to the transmon through a resonator in the dispersive regime, constituting a carefully engineered environmental spectrum for the transmon. Using single-shot readout, we observe excitation stabilization times down to roughly $500$ ns, a $20$-fold speedup with QCR and a simultaneous two-tone drive addressing the $\text{\$e\$-\$f\$}$ and $\text{\$f0\$-\$g1\$}$ transitions of the system. Our results are obtained at a $\text{\$48\$-\$mk\$}$ fridge temperature and without postselection, fully capturing the advantage of the protocol for the short-time dynamics and the drive-induced detrimental asymptotic behavior in the presence of relatively hot other baths of the transmon. We validate our results with a detailed Liouvillian model truncated up to the three-excitation subspace, from which we estimate the performance of the protocol in optimized scenarios, such as cold transmon baths and fine-tuned driving frequencies. These results pave the way for optimized reset of quantum-electric devices using engineered environments and for dissipation-engineered state preparation.  
Comments : 13 pages, 5 figures  
Date : 26 January 2024  
PDF_URL : https://arxiv.org/pdf/2401.14912  
-----  
#7  
-----  
Document ID : 90  
Title : Benchmarking Bayesian quantum estimation  
Authors : Valeria Cimini, Emanuele Polino, Mauro Valeri, Nicolò Spagnolo, Fabio Sciarrino  
Subjects : Quantum Physics  
Abstract : The quest for precision in parameter estimation is a fundamental task in different scientific areas. The relevance of this problem thus provided the motivation to develop methods for the application of quantum resources to estimation protocols. Within this context, Bayesian estimation offers a complete framework for optimal quantum metrology techniques, such as adaptive protocols. However, the use of the Bayesian approach requires extensive computational resources, especially in the multiparameter estimations that represent the typical operational scenario for quantum sensors. Hence, the requirement to characterize protocols implementing Bayesian estimations can become a significant challenge. This work focuses on the crucial task of robustly benchmarking the performances of these protocols in both single and multiple-parameter scenarios. By comparing different figures of merits, evidence is provided in favor of using the median of the quadratic error in the estimations in order to mitigate spurious effects due to the numerical discretization of the parameter space, the presence of limited data, and numerical instabilities. These results, providing a robust and reliable characterization of Bayesian protocols, find natural applications to practical problems within the quantum estimation framework.  
Comments :  
Date : 26 January 2024  
PDF_URL : https://arxiv.org/pdf/2401.14900
```

Εικόνα 5.Δ.1 Αποτελέσματα αναζήτησης ερωτήματος για αξιολόγηση

- Πληροφοριακή ανάγκη → Νανοτεχνολογία
- Ερώτημα χρήστη → quantum and computer
- Συναφείς κείμενα ως προς το ερώτημα → 4, 7, 73, 84, 90
- Συναφείς κείμενα ως προς την πληροφοριακή ανάγκη → 4, 84

$$P = \frac{\#\text{συναφή ανακτηθέντα κείμενα}}{\#\text{ανακτηθέντα κείμενα}} \rightarrow P = \frac{2}{7} \rightarrow P = 30\%$$

$$R = \frac{\#\text{συναφή ανακτηθέντα κείμενα}}{\#\text{συναφή κείμενα}} \rightarrow R = \frac{2}{5} \rightarrow R = 40\%$$

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

$$\frac{1}{F} = \frac{1}{2} \times \left(\frac{1}{P} + \frac{1}{R} \right) \rightarrow \frac{1}{F} = \frac{1}{2} \times \left(\frac{1}{0.3} + \frac{1}{0.4} \right) \rightarrow F \cong 35\%$$

Με βάση τις προαναφερόμενες μετρήσεις αξιολόγησης, η μηχανή αναζήτησης έχει μέτρια ακρίβεια και ανάκληση, ενώ η αντιστρόφως αναλογική συνάφεια προσδιορίζει την σημασία της πληροφορίας του κάθε όρου. Η απόδοση της μηχανής είναι σε ικανοποιητικά πλαίσια, ωστόσο υπάρχει πρόοδος για βελτίωση.

5.ε. Ανάλυση και Βελτιώσεις

Η μηχανή αναζήτησης δεν δύναται να φιλτράρει πολλαπλούς συγγραφείς και αυτό είναι κάτι το οποίο θα διευκόλυνε ακόμα περισσότερο τον χρήστη. Το GUI επίσης είναι αρκετά απλό και θα μπορούσε να είναι πιο εύχρηστο και πιο εμφανισιακά ωραίο όσον αφορά το layout του. Σίγουρα και η αποδοτικότητα των αλγορίθμων δεν είναι η καλύτερη δυνατή σε όλους, αλλά δεν επηρεάζει αυτό σημαντικά την ακρίβεια πόσο μάλλον την χρήση της εφαρμογής.

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



Σας ευχαριστούμε για την προσοχή σας.

