Search & NLP Bootcamp

August 2017

Ashish Kaduskar & John Kuriakose

Infosys®

Agenda

- Scope and Background
- Scope & tasks
- Infrastructure
- Indicative Architecture
- Appendix
  - Screenshots

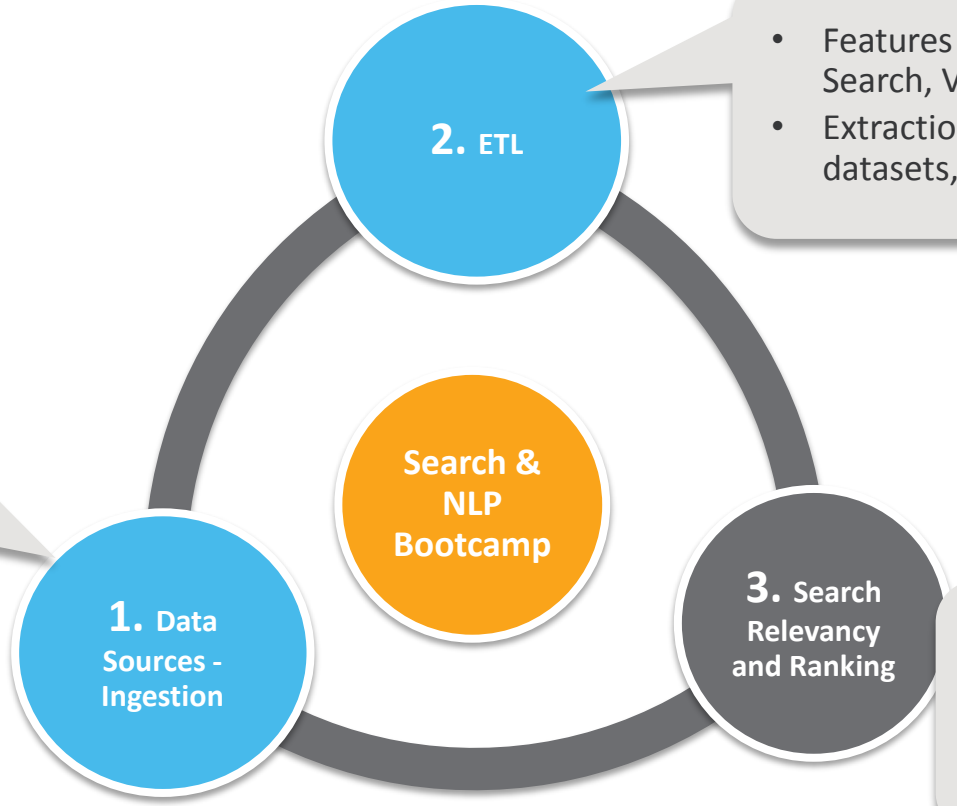
ARXIV

- The **arXiv** (pronounced "archive") is a repository of electronic preprints, known as e-prints, of scientific papers in the fields of mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance, which can be accessed online.
- **It** is an open-access database where you can read and download research papers from some quantitative scientific fields (used a lot for physics and astronomy). It is extremely useful for finding research papers on a given topic.
- https://arxiv.org/
- ARXIV provides an API to access Metadata of their dataset - https://arxiv.org/help/oa/index. Read about the protocol OAI-PMH
- ARXIV provides a dump of the dataset pdf documents and Latex source files on Amazon S3. The pdf documents are created from the Latex files. https://arxiv.org/help/bulk_data_s3

Scope

- Elsevier is looking for a strategic partner who has the expertise and scale to provide the required technology leadership and implementation support for DataSearch.

- Ingest a fragment of ARXIV pdf documents and process them in Spark
- Index documents into a SOLR Cloud Server
- Build and tune a search system over Research Data



- Features in scope – Enrichment(NLP), Indexing, Search, Visualization, Tools
- Extraction of granular information assets – datasets, figures and tables.

- Measure Precision and NDCG for a set of 10 test queries .
- Build an end-to-end Applications for creating a GOLD set (Judgment data) of results for each of the test queries

Bootcamp Tasks

Ingestion

- Build a harvester that can access and download ARXIV Metadata using the Arxiv OAI-PMH endpoint
- Write the collected Files on HDFS

Phase 2

- Access the ARXIV Bulk data on Amazon S3 and download PDF files with full text . (Phase 1 was only metadata).
- Extract Text with section info from PDF and write to HDFS
- Modify earlier Spark job to process ARXIV full Text
- Change SOLR schema as required

ETL

- Design suitable schema on SOLR for Metadata only Index
- Build a Spark Job to process these files and Index to SOLR
- Build and Tune Search

Phase 2 Advanced Option (optional)

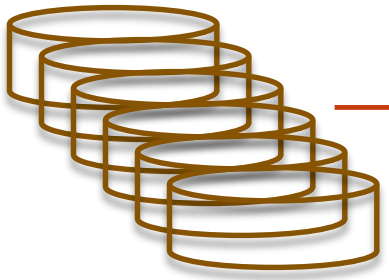
- Integrate NLP4J and lemmatize data and index lemmatized data
- Learn how to get Phrases from textual data and store Phrases into REDIS

Measure and Improve Relevance

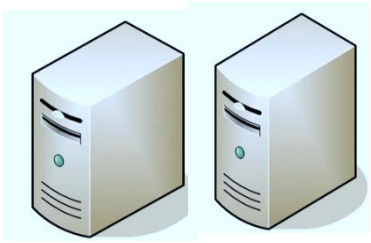
- Build end-to-end system for Search Relevance Judgment
- Select list of Test Queries – final list to have 10 test queries
- Build Judgment data for all test queries using the judgment system
- Measure precision and NDCG and show in Judgment UI
- Improve Search relevance and show improvement in metrics
- **Goal is to achieve 6 / 10 for precision**

Infosys®

Infrastructure Setup



ARXIV – Bulk Data Access
+ OAI



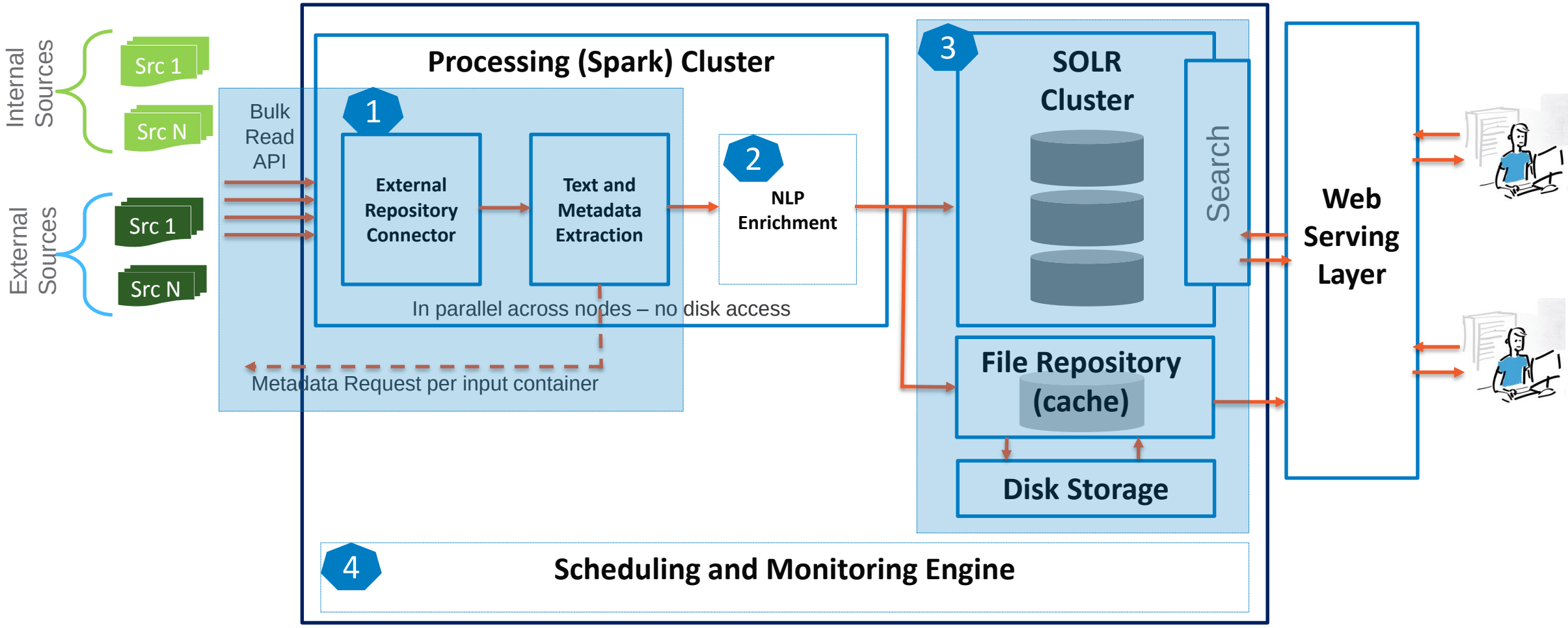
Azure Cloud Instance
Spark, Hadoop, Solr Cloud,
Tomcat, NLP

- Document type PDF and LaTeX.
- Arxiv API for bulk data access (in AWS S3) and an OAI API for Metadata
- Data set for Prototype 6 months data
- Total Research Papers (containers) = 1.2 million
- Total data size for is 890GB (approx.)

- ❑ The team will have access to a shared 2 node cluster in our development environment. (Platform Team)
- ❑ Spark, Hadoop setup for ingestion, processing and indexing. - To be done
- ❑ Solr Cloud setup for indexing and search. – To be done
- ❑ UI and search backend deployed on Tomcat (web server). - To be done

Infosys®

Indicative High Level Architecture



Architectural and Infrastructure Drivers

- Size of Input Data and Size of Index
- Processing Time for ingestion, extraction and enrichment

Infosys®

Screenshots of system you are expected to build

Key Technical considerations (indicative – not in scope)

Ingestion

1

- Ingest directly from External APIs or Stores (Amazon S3)
- Decompress (tar, gz, tar.gz etc) and process as Input Stream without first copying data to disk in cluster.
- Diskless approach is much faster and does not unnecessarily require disk space for maintaining a copy of raw data.
- Ingest in parallel, with file / folder level monitoring
- Support for multiple document formats (PDF, Latex, HTML) and external adaptors. (Amazon S3, FTP, URL)
- Delta Ingestion and Indexing framework with specific adaptors for each data source

Enrichment

2

- Run NLP Enrichment in Parallel
- Phrase Chunking, Named Entities, Ngrams, Query Expansion for Units of Measure.
- Extract Text and Metadata - find mentions and references to information assets e.g. Tables, Figures and Datasets.
- Scalability and performance is achieved with Spark's in memory processing .
- Leverage Spark for Automatic recovery in case of cluster node failure.
- Number of Nodes used for ingestion and enrichment is based on data size and the time taken. Dynamically add / remove nodes to manage performance.
- Spot nodes are ideally suited for this task since no permanent storage in Spark Cluster.

Store & Index - SOLR

3

- Enriched Text + Metadata flows to SOLR Cloud Cluster.
- SOLR Shards represents parallelism for Indexing and Search. Replication enabled for high availability and load balancing.
- Dedicated Nodes are used for SOLR Cluster.
- Clear separation of Clusters for enrichment processing and store (Index) allows to use varying QoS of EMR Nodes

Scheduling and Monitoring

4

- Workflow Management and Orchestration
- Monitoring job execution,
- Automated triggering for incremental data ingestion and Solr index updates

Infosys®

datasearchdemo.elsevier.com/indexed#/search/Effective%20field%20theory%20Framework?source=ARXIV

ASUS Login | 7 - 5 - Other Sentimer | deeplearning4nlp-tut | Infosys Information Pl | Zhiyuan (Brett) Chen's | Text & Data Mining b | TIME | smartdataweek.com/c | Other book

taSearch | Effective field theory Framework | Effective field theory Framework 1 of 2

Results reset

s

e (384020)

ar Data (112594)

ces

(402140)

1990-01-14

2016-02-18

402140 Results

Dark matter at the LHC: EFTs and gauge invariance
Bell, Nicole F., Cai, Yi, Dent, James B., Leane, Rebecca K. & Weiler, Thomas J. - 2015-03-26
Effective field theory (EFT) formulations of dark matter interactions have proven to be a convenient and popular way to quantify LHC bounds on dark matter. However, some of the non-renormalizable EFT operators considered do not respect the gauge symmetries of the Standard Model. We carefully discuss under what circumstances such operators can arise, and outline potential issues in their interpretation and application.
IMAGE 4

No-core shell model in an effective-field-theory framework
Stetcu, I., Barrett, B. R. & van Kolck, U. - 2006-09-11
We present a new approach to the construction of effective interactions suitable for many-body calculations by means of the no-core shell model (NCSM). We consider an effective field theory (EFT) with only nucleon fields directly in the NCSM model spaces. In leading order, we obtain the strengths of the three contact terms from the condition that in each model space the experimental ground-state energies of 2H, 3H and 4He be exactly reproduced. The first (0^+;0) excited state of 4He and the ground state of 6Li are then obtained by
IMAGE 3

Three and Four Harmonically Trapped Particles in an Effective Field Theory Framework
Rotureau, J., Stetcu, I., Barrett, B. R., Birse, M. C. & van Kolck, U. - 2010-08-26
We study systems of few two-component fermions interacting via short-range interactions within a harmonic-oscillator trap. The dominant interactions, which are two-body, are organized according to the number of derivatives and defined in a two-body truncated model space made from a bound-state basis. Leading-order (LO) interactions are solved for exactly using the formalism of the No-Core Shell Model, whereas corrections are treated as many-body perturbations. We show explicitly that next-to-LO and next-to-next-to-LO
IMAGE 10

NLO Higgs Effective Field Theory and kappa-framework
Ghezzi, Margherita, Gomez-Ambrosio, Raquel, Passarino, Giampiero & Uccirati, Sandro - 2015-05-14
A consistent framework for studying Standard Model deviations is developed. It assumes that New Physics becomes relevant at some scale beyond the present experimental reach and uses the Effective Field Theory approach by adding higher-dimensional operators to the Standard Model Lagrangian and by computing relevant processes at the next-to-leading order, extending the original kappa-framework.
IMAGE TABULAR DATA 4

Baryon chiral perturbation theory
Scherer, Stefan - 2011-12-23
We provide a short introduction to the one-nucleon sector of chiral perturbation theory and address the issue of power counting and renormalization. We discuss the infrared regularization and the extended on-mass-shell scheme. Both allow for the inclusion of further degrees of freedom beyond pions and nucleons and the application to higher-loop calculations. As applications we consider the chiral expansion of the nucleon mass to order ${\cal O}(q^6)$ and the inclusion of vector and axial-vector mesons in the calculation of
IMAGE 8 TABULAR DATA

Chiral Perturbation Theory: Introduction and Recent Results in the One-Nucleon Sector
Scherer, Stefan - 2009-08-24
We provide an introduction to the basic concepts of chiral perturbation theory and discuss some recent developments in the manifestly Lorentz-invariant formulation of the one-nucleon sector.

Results from DataSearch
Only 3 Docs in the Top 10 - Docs #2 #4 and #5 contain the User's Search Query

datasearchdemo.elsevier.com/indexed#/search/Thermal%20Dark%20Matter%20Particles

ASUS Login | 7 - 5 - Other Sentimer | deeplearning4nlp-tut | Infosys Information Pl | Zhiyuan (Brett) Chen's | Text & Data Mining b | TIME | smartdataweek.com/c | Other book

taSearch

Thermal Dark Matter Particles

dark matter particles 1 of 11

Results reset

272221 Results

s

e (254390)

ar Data (75208)

ment (3866)

Data (812)

(152)

et (83)

s (69)

tical Data (20)

are (1)

ces

(211449)

ceDirect (42856)

ed Central (14893)

oElectro (1327)

moML at NIST TRC (411)

B (378)

Semi-annihilation of Dark Matter

D'Eramo, Francesco - 2011-01-27

The semi-annihilation reaction takes the schematic form psi_i psi_j -> psi_k phi, where psi_i are stable dark matter particles and phi is an unstable state. Such reactions are allowed when dark matter is stabilized by a larger symmetry than just Z_2. They lead to non-trivial dark matter dynamics in the early universe, and the thermal production of the relic particles can be completely controlled by semi-annihilations. This process might also take place today in the Milky Way, enriching the (semi-)annihilation final state spectrum observed in

IMAGE 4 TABULAR DATA

Limits on MeV Dark Matter from the Effective Number of Neutrinos

Ho, Chiu Man & Scherrer, Robert J. - 2012-12-17

Thermal dark matter that couples more strongly to electrons and photons than to neutrinos will heat the electron-photon plasma relative to the neutrino background if it becomes nonrelativistic after the neutrinos decouple from the thermal background. This results in a reduction in N_eff below the standard-model value, a result strongly disfavored by current CMB observations. Taking conservative lower bounds on N_eff and on the decoupling temperature of the neutrinos, we derive a bound on the dark matter particle mass of m_\chi > 3-

IMAGE

Light Dark Matter and Dark Radiation

Heo, Jae Ho & Kim, C. S. - 2016-02-16

Light dark-matter ($M\leq20$ MeV) particles freeze out after neutrino decoupling. If the dark-matter particle couples to a neutrino or an electromagnetic plasma, the late time entropy production from dark-matter annihilation can change the neutrino-to-photon temperature ratio, and equally the effective number of neutrinos N_{eff}. We study the non-equilibrium effects of dark-matter annihilation on the N_{eff} and the effects by using a thermal equilibrium approximation. Both results are constrained with Planck observations. We

IMAGE 3 TABULAR DATA 2

Dark Matter and Dark Radiation

Ackerman, Lotty, Buckley, Matthew R., Carroll, Sean M. & Kamionkowski, Marc - 2008-12-15

We explore the feasibility and astrophysical consequences of a new long-range U(1) gauge field ("dark electromagnetism") that couples only to dark matter, not to the Standard Model. The dark matter consists of an equal number of positive and negative charges under the new force, but annihilations are suppressed if the dark matter mass is sufficiently high and the dark fine-structure constant $\hat\alpha$ is sufficiently small. The correct relic abundance can be obtained if the dark matter also couples to the conventional weak interactions,

IMAGE 0

Dark Matter Stability without New Symmetries

Cata, Oscar & Ibarra, Alejandro - 2014-08-28

The stability of dark matter is normally achieved by imposing extra symmetries beyond those of the Standard Model of Particle Physics. In this paper we present a framework where the dark matter stability emerges as a consequence of the Standard Model symmetries. The dark m[Preview Data]symmetric tensor field (analogous to the one used for spin-1 mesons in QCD), singlet under the Standard Model gauge group. The Lagrangian possesses an accidental Z_2 symmetry which makes the dark matter stable on cosmological time

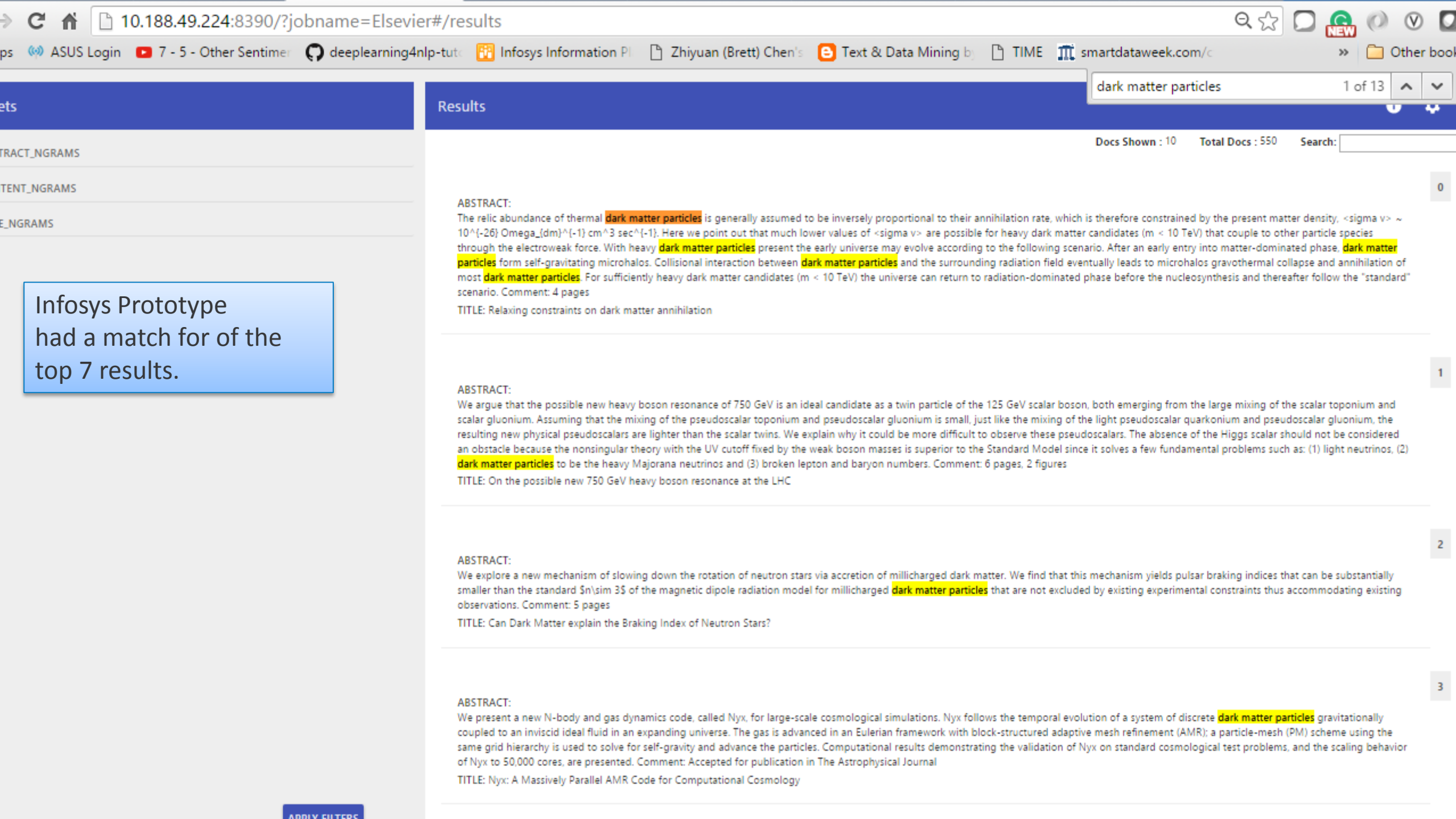
IMAGE 5

Dark Matter Candidates from Particle Physics and Methods of Detection

Feng, Jonathan L. - 2010-04-09

The identity of dark matter is a question of central importance in both astrophysics and particle physics. In the past, the leading particle candidates were cold and collisionless, and typically predicted missing energy signals at particle colliders. However, recent progress has greatly expanded the list of well-motivated candidates and the possible signatures of dark

Results from DataSearch
Results match 'dark matter' rather than 'dark matter particles'
Only 1 Doc in the Top 10 - Doc #1 contains query


10.188.49.224:8390/?jobname=Elsevier#/results
ASUS Login
7 - 5 - Other Sentimer
deeplearning4nlp-tut
Infosys Information Pl
Zhiyuan (Brett) Chen's
Text & Data Mining b
TIME
smartdataweek.com/c
Other book
ets
TRACT_NGRAMS
TENT_NGRAMS
E_NGRAMS
Infosys Prototype had a match for of the top 7 results.
APPLY FILTERS
Results
dark matter particles
1 of 13
Docs Shown : 10
Total Docs : 550
Search:
0
ABSTRACT:
The relic abundance of thermal dark matter particles is generally assumed to be inversely proportional to their annihilation rate, which is therefore constrained by the present matter density, <sigma v> ~ 10^{-26} Omega_{dm}^{-1} cm^3 sec^{-1}. Here we point out that much lower values of <sigma v> are possible for heavy dark matter candidates (m < 10 TeV) that couple to other particle species through the electroweak force. With heavy dark matter particles present the early universe may evolve according to the following scenario. After an early entry into matter-dominated phase, dark matter particles form self-gravitating microhalos. Collisional interaction between dark matter particles and the surrounding radiation field eventually leads to microhalos gravothermal collapse and annihilation of most dark matter particles. For sufficiently heavy dark matter candidates (m < 10 TeV) the universe can return to radiation-dominated phase before the nucleosynthesis and thereafter follow the "standard" scenario. Comment: 4 pages
TITLE: Relaxing constraints on dark matter annihilation
1
ABSTRACT:
We argue that the possible new heavy boson resonance of 750 GeV is an ideal candidate as a twin particle of the 125 GeV scalar boson, both emerging from the large mixing of the scalar toponium and scalar gluonium. Assuming that the mixing of the pseudoscalar toponium and pseudoscalar gluonium is small, just like the mixing of the light pseudoscalar quarkonium and pseudoscalar gluonium, the resulting new physical pseudoscalars are lighter than the scalar twins. We explain why it could be more difficult to observe these pseudoscalars. The absence of the Higgs scalar should not be considered an obstacle because the nonsingular theory with the UV cutoff fixed by the weak boson masses is superior to the Standard Model since it solves a few fundamental problems such as: (1) light neutrinos, (2) dark matter particles to be the heavy Majorana neutrinos and (3) broken lepton and baryon numbers. Comment: 6 pages, 2 figures
TITLE: On the possible new 750 GeV heavy boson resonance at the LHC
2
ABSTRACT:
We explore a new mechanism of slowing down the rotation of neutron stars via accretion of millicharged dark matter. We find that this mechanism yields pulsar braking indices that can be substantially smaller than the standard $n\sim 3$ of the magnetic dipole radiation model for millicharged dark matter particles that are not excluded by existing experimental constraints thus accommodating existing observations. Comment: 5 pages
TITLE: Can Dark Matter explain the Braking Index of Neutron Stars?
3
ABSTRACT:
We present a new N-body and gas dynamics code, called Nyx, for large-scale cosmological simulations. Nyx follows the temporal evolution of a system of discrete dark matter particles gravitationally coupled to an inviscid ideal fluid in an expanding universe. The gas is advanced in an Eulerian framework with block-structured adaptive mesh refinement (AMR); a particle-mesh (PM) scheme using the same grid hierarchy is used to solve for self-gravity and advance the particles. Computational results demonstrating the validation of Nyx on standard cosmological test problems, and the scaling behavior of Nyx to 50,000 cores, are presented. Comment: Accepted for publication in The Astrophysical Journal
TITLE: Nyx: A Massively Parallel AMR Code for Computational Cosmology

Thank you