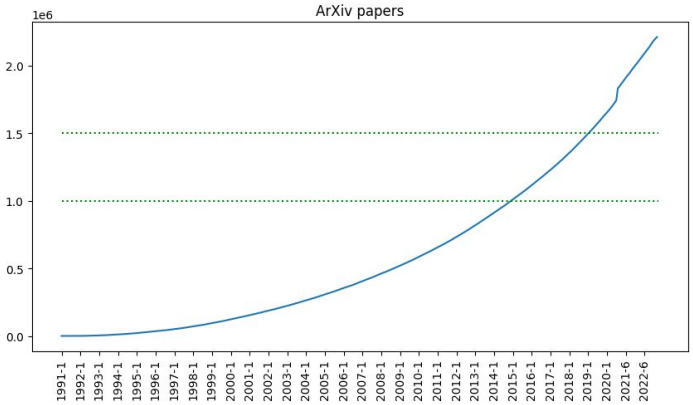
Project DeepDipper

NLP toolkits to master your exploration of research papers

March 2023



Observation 1 - *begun on August 14,* ***1991****, arXiv.org contains today* ***over 2.2 million papers.***
Observation 2 - *in* ***deep learning*** *and* ***AI****,* what is true *now* is **likely to be outdated in 6 months**.

🚨 Urgency for knowledge parsing and recommendations tools at scale 🚨

arXiv is an ***open-access repository of electronic preprints*** *(known as e-prints) approved for posting after moderation, but not full peer reviewed. In many fields of* ***mathematics****,* ***physics*** *and* ***computer sciences****, almost all scientific papers are* ***self-archived*** *on the arXiv repository* ***before publication*** *in a peer-reviewed journal.*

Project DeepDipper - Building an all-in-one NLP toolkit to navigate and extract info from fields of research papers

🤲 ***Project basis*** → last three decades of research papers in data science and related fields in the broadest sense.

🎯 ***Goal*** → a minimally working interactive website, to be potentially polished later if affinities.

⛳ ***Work objectives:***

A) **an interactive analytics dashboard**: for an overview of areas of research key info and characteristics

B) a set of functionalities available for users, leveraging **multiple NLP algorithms**

Data available

Dataset of >2.2m research papers in areas related (more or less) to data science; arXiv consists of scientific papers, categorized in the fields of these groups:

- Computer Science
- Economics
- Electrical Engineering and Systems Science
- Mathematics
- Physics
- Quantitative Biology
- Quantitative Finance
- Statistics

Dataset of ~3000 research papers focused on all machine learning areas, scraped from website of the Journal of Machine Learning Research (JMLR).

We can focus more closely on specific research areas and do more advanced analysis by extracting text and graphic data from pdf links (with OCRs).

— Available features in arXiv dataset —

ID | Title | Datetime | Authors | Submitter

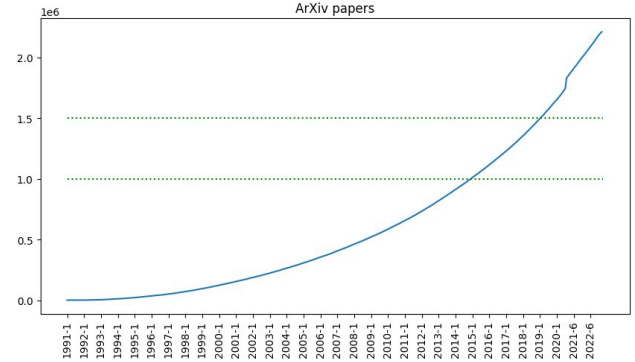
PDF url | Abstract | Categories | Journal ref

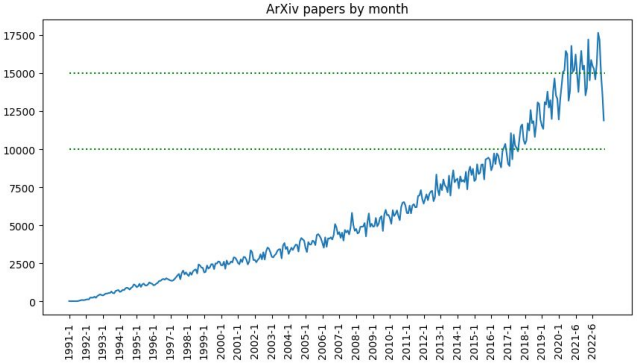
— Other available data —

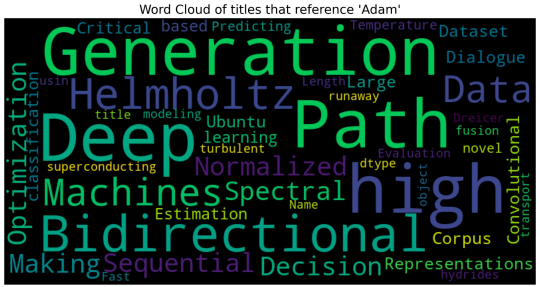
For arXiv: Citations | Taxonomy | License

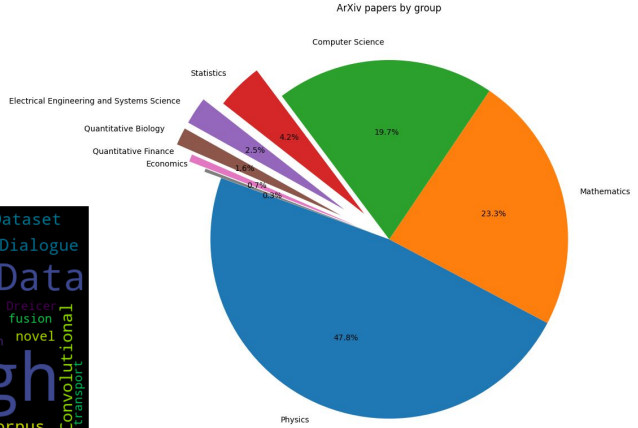
\+ JMLR dataset (with main features & urls)

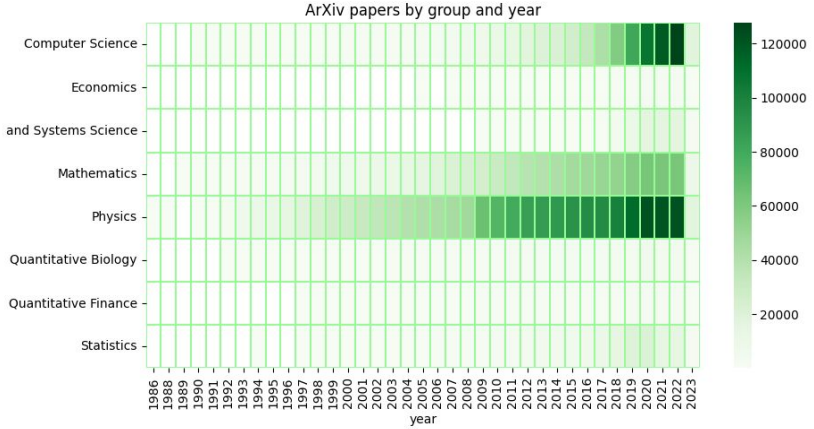
DeepDipper's Interactive analytics dashboard











DeepDipper's NLP-based functionalities

Module 1				Module 2	Module 3
Core component *difficulty:* 🟡	**Core component** *difficulty:* 🟡	**Optional component** *difficulty:* 🟢	**Optional component** *difficulty:* 🔴	**Core component** *difficulty:* 🟢	**Optional component** *difficulty:* 🟢
Recommendation of papers	**Filtering/searching topics in papers**	**Multi-classification of papers**	**Citation network analysis**	**Summarisation of abstracts**	**Translation of titles and abstracts**
"Recommend top N papers based on user-selected papers"	"Find top N papers based on user-written sentence/tags"	"Add 2nd-ary categories for papers based on abstract"	"Analyze papers' ties based on citations and authors"	"Summarize this paper's abstract into a one-liner"	"Translate this paper's abstract and title into French"
K-Nearest Neighbors Algorithm (KNN) *(supervised model)*	Tfidf vectors, Bert, or Latent Dirichlet Allocation (LDA) *(unsupervised model)*	'Roberta' model, or zero shot with Hugging Face *(unsupervised model)*	Graph Convolutional Networks (GCNs) *(unsupervised model)*	PegasusX5, Bart or simple T5 model *(unsupervised &/or transfer learning)*	Seq2SeqModel "MarianMT" *(transfer learning)*
Trained with: 'title', 'authors', 'year', 'abstract' Input: <user prompt paper ID(s)>	Trained with: 'title', 'authors', 'year', 'abstract' Input: <user prompt string word(s)>	Trained with: 'title', 'abstract', 'id', 'year', 'categories'	Trained with: 'title', 'authors', 'year', 'id', 'abstract', 'citations'	Trained with: 'title', 'abstract' Input: <user prompt integer index>	Trained with: 'title', 'abstract' Input: <user prompt integer index>

A project to explore a panel of advanced NLP techniques, expand your meta comprehension of (data) sciences, and build a tool potentially useful to the broader community

Looking for all kinds of brains and souls to join the project! Let’s work together, especially if you:

- Have profound experience in reading, collecting and sorting through stuff in general.
- Are curious to get an eagle-eye view of the researches done in areas related to data science.
- Are eager to explore various NLP algorithms, and their applications at scale and in production.
- Feel strongly about the need to democratize tools that help in parsing available knowledge/content.

```
'astro-ph': 'Astrophysics',
'astro-ph.CO': 'Cosmology and Nongalactic Astrophysics',
'astro-ph.EP': 'Earth and Planetary Astrophysics',
'astro-ph.GA': 'Astrophysics of Galaxies',
'astro-ph.HE': 'High Energy Astrophysical Phenomena',
'astro-ph.IM': 'Instrumentation and Methods for Astrophysics
'astro-ph.SR': 'Solar and Stellar Astrophysics',
'cond-mat.dis-nn': 'Disordered Systems and Neural Networks',
'cond-mat.mes-hall': 'Mesoscale and Nanoscale Physics',
'cond-mat.mtrl-sci': 'Materials Science',
'cond-mat.other': 'Other Condensed Matter',
'cond-mat.quant-gas': 'Quantum Gases',
'cond-mat.soft': 'Soft Condensed Matter',
'cond-mat.stat-mech': 'Statistical Mechanics',
'cond-mat.str-el': 'Strongly Correlated Electrons',
'cond-mat.supr-con': 'Superconductivity',
'cs.AI': 'Artificial Intelligence',
'cs.AR': 'Hardware Architecture',
'cs.CC': 'Computational Complexity',
'cs.CE': 'Computational Engineering, Finance, and Science',
'cs.CG': 'Computational Geometry',
'cs.CL': 'Computation and Language',
'cs.CR': 'Cryptography and Security',
'cs.CV': 'Computer Vision and Pattern Recognition',
'cs.CY': 'Computers and Society',
'cs.DB': 'Databases',
'cs.DC': 'Distributed, Parallel, and Cluster Computing',
'cs.DL': 'Digital Libraries',
'cs.DM': 'Discrete Mathematics',
'cs.DS': 'Data Structures and Algorithms',
'cs.ET': 'Emerging Technologies',
'cs.RO': 'Robotics',
'cs.SC': 'Symbolic Computation',
'cs.SD': 'Sound',
'cs.SE': 'Software Engineering',
'cs.SI': 'Social and Information Networks',
'cs.SY': 'Systems and Control',
'econ.EM': 'Econometrics',
'eess.AS': 'Audio and Speech Processing',
'eess.IV': 'Image and Video Processing',
'eess.SP': 'Signal Processing',
'gr-qc': 'General Relativity and Quantum Cosmology',
'hep-ex': 'High Energy Physics - Experiment',
'hep-lat': 'High Energy Physics - Lattice',
'hep-ph': 'High Energy Physics - Phenomenology',
'hep-th': 'High Energy Physics - Theory',
'math.AC': 'Commutative Algebra',
'math.AG': 'Algebraic Geometry',
'math.AP': 'Analysis of PDEs',
'math.AT': 'Algebraic Topology',
'math.CA': 'Classical Analysis and ODEs',
'math.CO': 'Combinatorics',
'math.CT': 'Category Theory',
'math.CV': 'Complex Variables',
'math.DG': 'Differential Geometry',
'math.DS': 'Dynamical Systems',
'math.FA': 'Functional Analysis',
'math.GM': 'General Mathematics',
'math.GN': 'General Topology',
'math.GR': 'Group Theory',
```