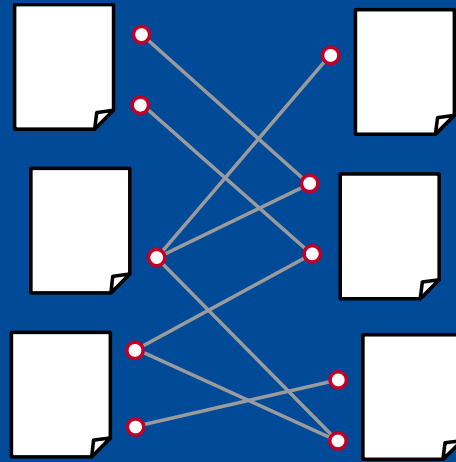


Bibliometric-Enhanced arXiv

A Data Set for Paper-Based and Citation-Based Tasks



Albert-Ludwigs-Universität Freiburg

Tarek Saier, Michael Färber
BIR Workshop 2019



**UNI
FREIBURG**

Background



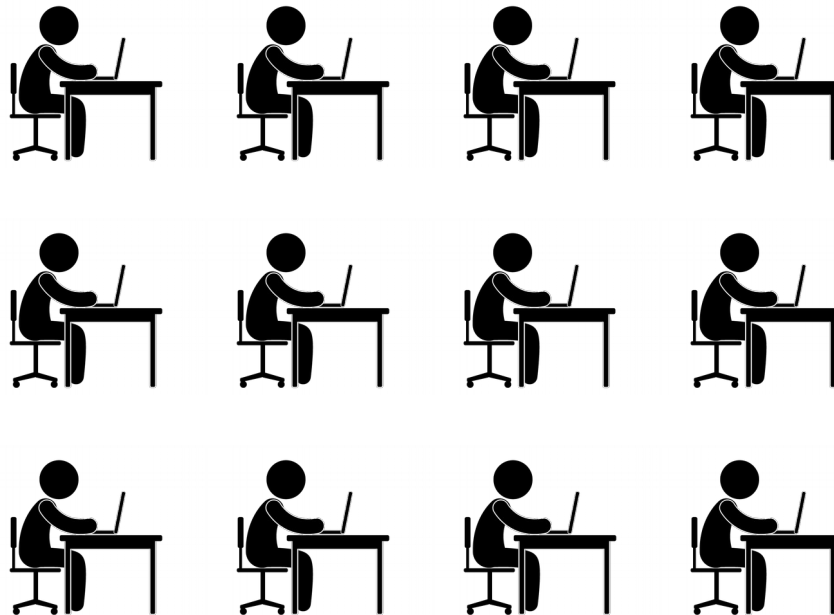
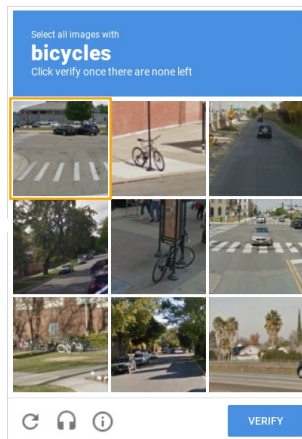
- Mining scholarly discourse
 - Paper recommendation
 - Citation recommendation
 - Citation context-based document summarization
 - Detection of bias
 - Detection of plagiarism
 - ...



Background



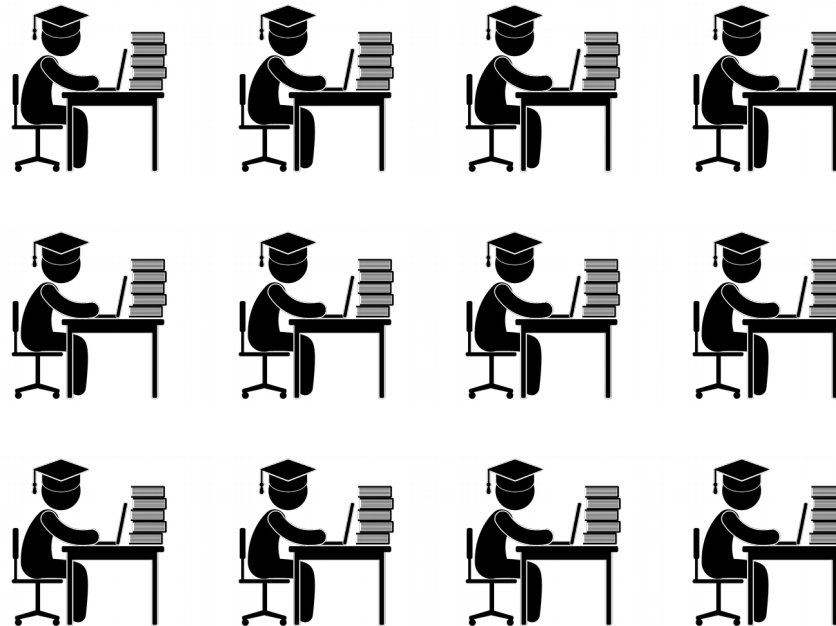
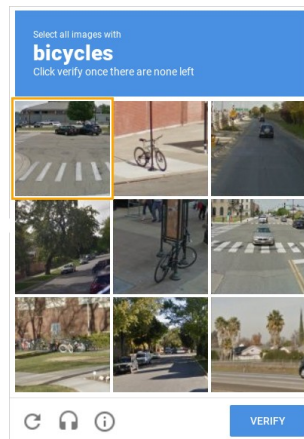
- Crowdsourcing image annotations



Background



■ Crowdsourcing ...



Background



■ Crowdsourcing citations



Background



- Lots of valuable information to be harvested

but ...

- Utilization requires
 - Machine readable formats
 - Sufficient amounts of data
 - Clean data
 - ...

Existing Data Sets



- PDF
 - Scholarly
 - ACL-AAN
 - ACL-ARC*
- JATS XML
 - PMC-OAS
 - PLOS
- Extracted plain text
 - CiteSeer^x
 - arXiv CS

Existing Data Sets



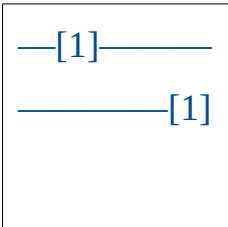
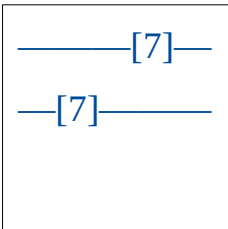
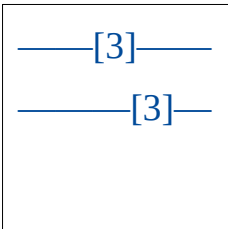
- PDF
 - Scholarly 200k
 - ACL-AAN 18k
 - ACL-ARC* 11k
- JATS XML
 - PMC-OAS 2M
 - PLOS 200k
- Extracted plain text
 - CiteSeer^x 5M
 - arXiv CS 90k

Existing Data Sets



- PDF
 - Scholarly 200k -
 - ACL-AAN 18k -
 - ACL-ARC* 11k -
- JATS XML
 - PMC-OAS 2M PubMed, MEDLINE, ...
 - PLOS 200k -
- Extracted plain text
 - CiteSeer^x 5M internal
 - arXiv CS 90k DBLP

Existing Data Sets



describing: citing documents

Existing Data Sets



—[3]—

—[3]—

[3] V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102

—[7]—

—[7]—

[7] V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.

—[1]—

—[1]—

[1] V. N. Şenoğuz and Q. Shafi, "Reheat temperature in super-Symmetric hybrid inflation models," Phys. Rev. D 71, 043514 (2005) [hep-ph/0412102].

describing: citing documents

Existing Data Sets



—[3]—

—[3]—

[3] V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102

—[7]—

—[7]—

[7] V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.

—[1]—

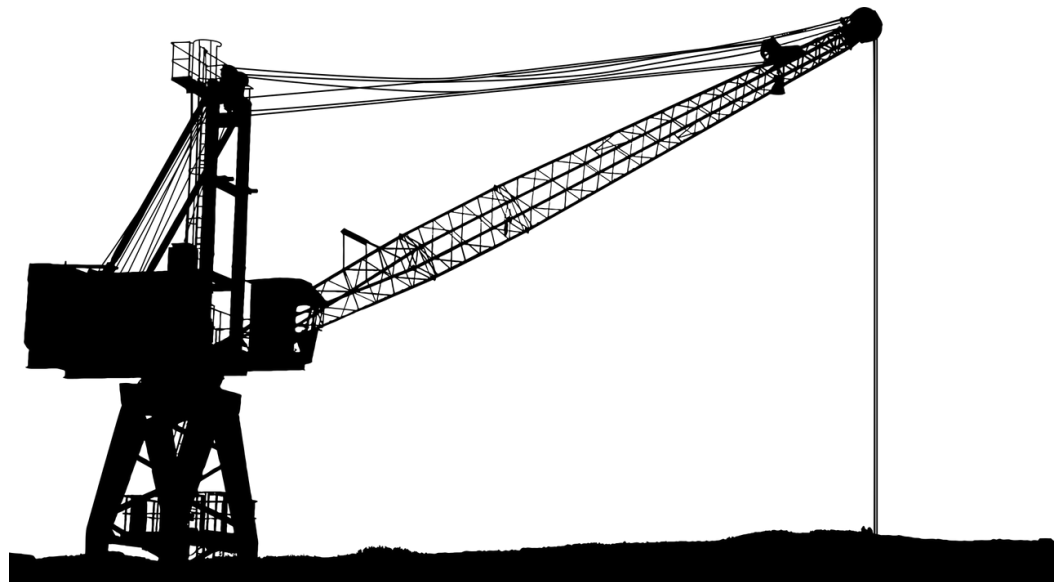
—[1]—

[1] V. N. Şenoğuz and Q. Shafi, "Reheat temperature in super-Symmetric hybrid inflation models," Phys. Rev. D 71, 043514 (2005) [hep-ph/0412102].



describing: citing documents + cited document

A New Dataset



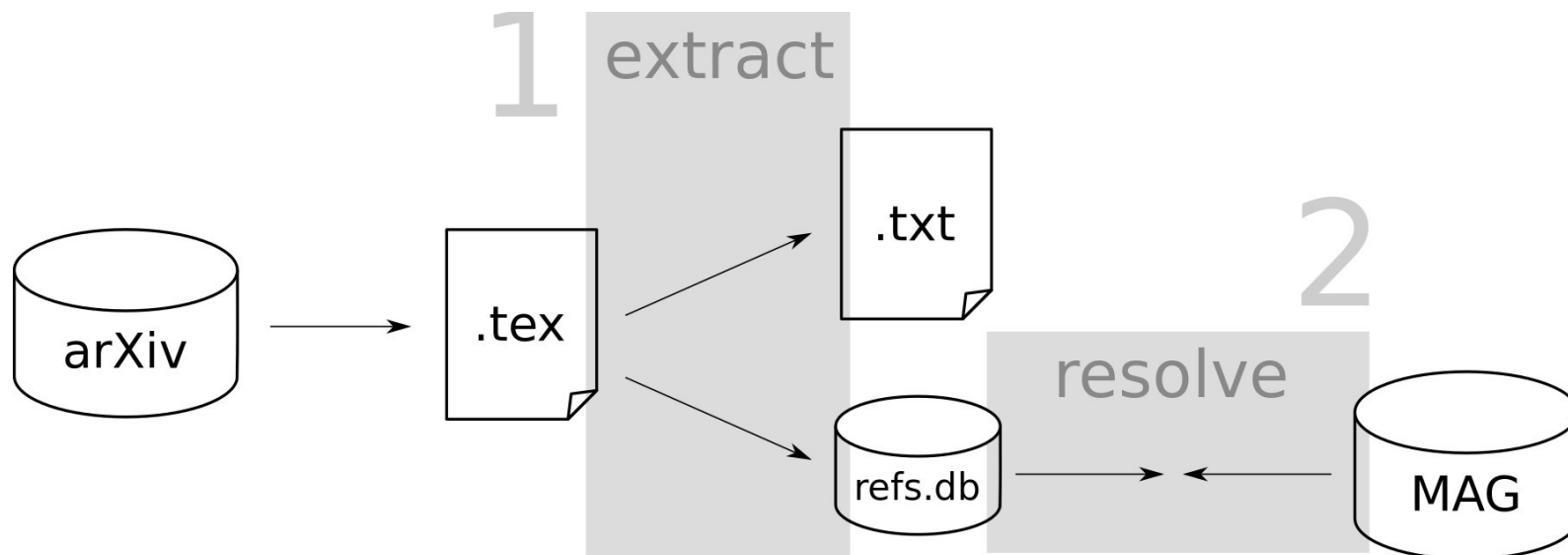
- **arXiv.org**
 - Operates since 1991
 - 1.5M papers
 - LaTeX sources of papers available

- **Microsoft Academic Graph (MAG)**
 - Automatically generated
 - Metadata on 213M papers

Creation Process



- Generation process



Creation Process



- Extract (93%)
 - flatten
 - convert to XML
 - extract text + references
- Resolve references (43%)
 - identify title (arXiv ID, DOI, Neural ParsCit)
 - match with MAG

Challenges



- Extraction
- Reference resolution

Challenges



```
@InProceedings{White2016,  
  author    = {White, Aaron Steven and Reisinger, Drew and Sakaguchi, Keisuke and Vieira,  
Tim and Zhang, Sheng and Rudinger, Rachel and Rawlins, Kyle and Van Durme, Benjamin},  
  title     = {{Universal Compositional Semantics on Universal Dependencies}},  
  booktitle = {Proceedings of the 2016 Conference on Empirical Methods in Natural Language  
Processing},  
  year      = {2016},  
  pages     = {1713--1723},  
  publisher = {Association for Computational Linguistics},  
  doi       = {10.18653/v1/D16-1177},  
  location  = {Austin, Texas},  
  url       = {http://aclweb.org/anthology/D16-1177},  
}
```

```
@InProceedings{Bollacker1998,  
  author    = {Bollacker, Kurt D. and Lawrence, Steve and Giles, C. Lee},  
  title     = {{CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identificati  
on of Interesting Publications}},  
  booktitle = {Proceedings of the Second International Conference on Autonomous Agents},  
  year      = {1998},  
  series    = {AGENTS '98},  
  pages     = {116--123},  
  address   = {New York, NY, USA},  
  publisher = {ACM},  
  acmid    = {280786},
```

1064,1

63%

Challenges



```
\begin{thebibliography}{8}

\bibitem{OMSpoluicao} World Health Organization, \textit{Pelo menos 2 milhões de pessoas m
orrem por ano no mundo por causa de água contaminada}. Brasília: Agência Brasil, 2011.
URL http://agenciabrasil.ebc.com.br/geral/noticia/2015-04/oms-estima-2-milhoes-de-mortes-por-comida-e-agua-contaminadas-todos-os-anos.

\bibitem{herreraplan} Município de Luruaco, \textit{Plan de desarrollo municipal de Luruaco
2012-2015}. Barranquilla: Editorial Universidad del Atlántico, 2012.

\bibitem{world1987guias} World Health Organization, \textit{Guías para la calidad del agua
potable}. Geneva: Ediciones de la OMS, 1988.

\bibitem{Servais} A. De Brauwere, N. K. Quattara, and P. Servais, ``Modeling fecal indicat
or bacteria concentrations in natural surface waters: A review,'' {\em Crit Rev Env Sci Te
c}, vol. 44, pp. 2380--2453, 2014.

\bibitem{LIU} W. -C. Liu, W. -T. Chan, and C. -C. Young, ``Modeling fecal coliform contami
nation in tidal Danshuei River estuarine system,'' {\em Sci Total Environ}, vol. 50, pp. 6
32--640, 2014.

\bibitem{romeiro} N. M. L. Romeiro, R. G. Castro, E. R. Cirilo, and P. L. Natti, ``Local c
alibration of coliforms parameters of water quality problem at Igapó I Lake, Londrina, Par
aná, Brazil,'' {\em Ecol Model}, vol. 222, pp. 1888--1896, 2011.
```



407,0-1

90%

Challenges



- “LaTeX sources”

```
\pdfoutput=1
\documentclass{article}
\usepackage[final]{pdfpages}
\begin{document}
\includepdf[pages=1-last]{rl-es.pdf}
\end{document}
```

Challenges



- newcommand for newcommand

```
\documentstyle[aps,preprint,tighten,floats,epsfig,amsfonts]{revtex}
\begin{document}
\newcommand{\nc}{\newcommand}
\nc{\be}{\begin{equation}}
\nc{\ee}{\end{equation}}
\nc{\bib}{\bibitem}

...

\begin{references}

%
\bibitem{Peskin:1995} M.E.\ Peskin and D.V.\ Schroeder, "An Introduction to
Quantum Field Theory," (Addison-Wesley, New York, 1995) 842 pp.
%
\bibitem{proc:latt97} Lattice 99, Nucl. Phys. (Proc.\ Suppl.) {\bf 83},
(2000).
%
\bib{mackenzie} G.P.\ Lepage and D.B.\ Mackenzie, Phys.\ Rev.\ D {\bf 48},
2250 (1993), hep-lat/9209022.
%
\bib{lepage} G.P.\ Lepage, ``Redesigning lattice QCD'',
hep-lat/9607076.
```

8,5

Top

Challenges



- Extraction
 - .bbl not .bib
 - Free spirited use of LaTeX
 - External packages for citations (e.g. natbib)
- Reference resolution

Challenges



- Only minimal information
 - V.Sauli, JHEP 02, 001 (2003).
- Formulas in titles
 - Aaij, Roel, et al. "Search for the $B^0_s \rightarrow \eta' \phi$ decay" Journal of High Energy Physics 2017.5 (2017): 158.

Challenges



- “hep-th.”

+

ic.microsoft.com/#/detail/2811252340

Microsoft Academic

hep-th.

hep-th.

Jul 2nd 2018, arXiv: Computation and Language

Yang-Hui He, Vishnu Jejjala, Brent D. Nelson

We apply techniques in natural language processing, computational linguistics, and machine-learning to investigate papers in arXiv: hep-ph, hep-lat, gr-qc, and math-ph. All of the titles of papers in each of these sections, from the inception of the sections, are treated as a corpus which we use to train the neural network Word2Vec. A comparative study of common n-grams, line overlaps, and similarities is carried out. We find notable scientific and sociological differences between the fields. In conjunction with the syntactic structure of the titles in different sub-fields of high energy and mathematical physics are sufficiently different to allow classification of formal versus phenomenological sections with 87.1% accuracy, and can perform a finer five-fold classification with 71.1% accuracy.

20 REFERENCES

7 CITATIONS*

RELATED PUBLICATIONS

Linked References

[Distributed Representations of Words and Phrases and their Compositional...](#)

Challenges



- “K. Kondo, hep-th/0303251.”

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

Neural ParsCit

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

- “hep th 0303251”

Neural ParsCit
normalize

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

- “hep th 0303251”
- “hep th”

Neural ParsCit
normalize
try substrings

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

- “hep th 0303251”
- “hep th”
- 2341106557 ✓

Neural ParsCit

normalize

try substrings

title match

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

- “hep th 0303251”
- “hep th”
- 2341106557 ✓
- Yang-Hui He, ✓
Vishnu Jejjala,
Brent D. Nelson

Neural ParsCit

normalize

try substrings

title match

author match

Challenges



- “K. Kondo, hep-th/0303251.”

author author title

- “hep th 0303251”
- “hep th”
- 2341106557 ✓
- Yang-Hui he, ✓
Vishnu Jejjala,
Brent D. Nelson

Neural ParsCit

normalize

try substrings

title match

author match

Challenges



- Extraction
 - .bbl not .bib
 - Free spirited use of LaTeX
 - External packages for citations (e.g. natbib)
- Reference resolution
 - Minimal information
 - Formulas in titles
 - MAG noise

- Structure
 - arXiv full texts (txt)
 - in-text annotated citations
 - resolved to MAG IDs
- Size
 - citing papers (full text): 1M
 - 663k phys / 237k math / 112k CS / 31k other
 - cited papers (pointer): 2.7M
 - resolved references: 16M
 - citation contexts: 29M

- Quality of reference resolution
 - Manually check 300
 - 3 errors → accuracy estimate: 96%

#		Document
1	matched	<i>"The Maunder Minimum"</i> (John A. Eddy; 1976)
	correct	<i>"The Maunder Minimum: A reappraisal"</i> (John A. Eddy; 1983)
2	matched	<i>"Support Vector Machines"</i> (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013)
	correct	<i>"1-norm Support Vector Machines"</i> (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003)
3	matched	<i>"The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water"</i> (David Chandler, David Limmer; 2013)
	correct	<i>"The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II"</i> (David T. Limmer, David Chandler; 2011)

■ Format

1412.3684.txt

[...] It has over 79 million images stored at the resolution of FORMULA . Each image is labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed-10a1cf89a2e2}} {{cite:298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexical database. [...]

refs.db

uuid	in_doc	mag_id	reference_string
9ad20b...	1412.3684	2081580037	George A. Miller (1995). Wo...
298db7...	1412.3684	2038721957	Christiane Fellbaum (1998),...

■ Format

MAG

paperid	originaltitle	publisher	
2081580037	WordNet : an electronic lexical database	MIT Press	
2038721957	WordNet: a lexical database for English	ACM	...

extracted_contexts.csv

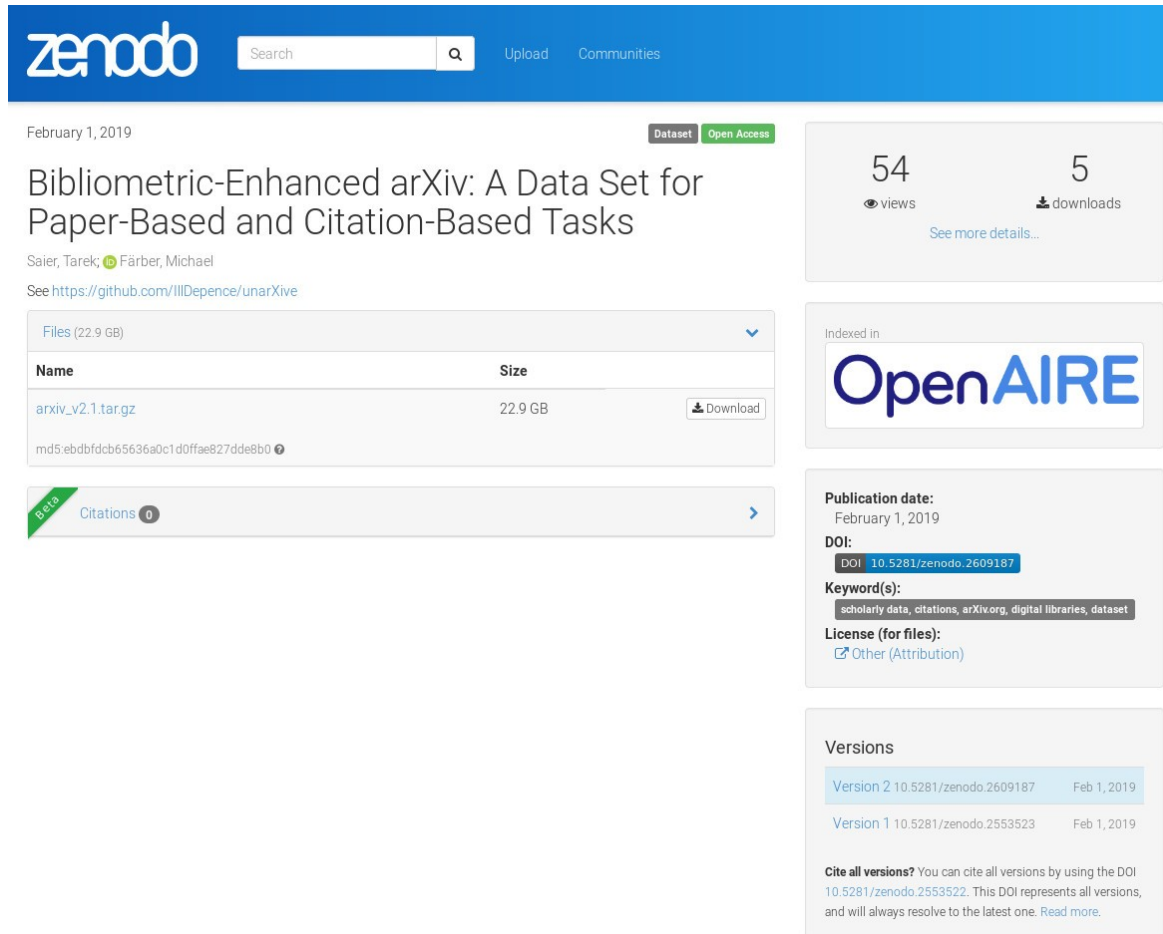
2038721957 | 2081580037 | 1412.3684 | It has over 79 million images stored at the resolution of FORMULA . Each image is labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet CIT MAINCIT lexical database. It has been noted that many of the labels are not reliable CIT .

- Context export script

```
~/arxiv_v2.1/code$ python3 extract_contexts.py  
usage: python3 extract_contexts.py </path/to/in/dir> [<db_uri>]
```

- context_size (in words/sentences)
- min_contexts
- min_citing_docs
- sample_size
- restrict_fos_citing

■ Data on Zenodo



The screenshot shows the Zenodo interface for a dataset. The header includes the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. The dataset title is 'Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks', dated February 1, 2019, and labeled as a 'Dataset' with 'Open Access'. It has 54 views and 5 downloads. The dataset is indexed in OpenAIRE. The file list shows 'arxiv_v2.1.tar.gz' (22.9 GB) with a download button. The citation count is 0. The publication date is February 1, 2019, and the DOI is 10.5281/zenodo.2609187. The keyword is 'scholarly data, citations, arXiv.org, digital libraries, dataset'. The license is 'Other (Attribution)'. The versions section shows two versions: Version 2 (10.5281/zenodo.2609187) and Version 1 (10.5281/zenodo.2553523), both dated February 1, 2019. A note at the bottom states that all versions can be cited using the DOI 10.5281/zenodo.2553522.

zenodo Search Upload Communities

February 1, 2019 Dataset Open Access

Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks

Saier, Tarek; Färber, Michael

See <https://github.com/IIIDepence/unarXiv>

Files (22.9 GB)

Name	Size	
arxiv_v2.1.tar.gz	22.9 GB	Download
md5:ebdbfdbc65636a0c1d0ffae827dde8b0		

Beta Citations 0

Indexed in OpenAIRE

Publication date: February 1, 2019

DOI: DOI 10.5281/zenodo.2609187

Keyword(s): scholarly data, citations, arXiv.org, digital libraries, dataset

License (for files): Other (Attribution)

Versions

Version	DOI	Date
Version 2	10.5281/zenodo.2609187	Feb 1, 2019
Version 1	10.5281/zenodo.2553523	Feb 1, 2019

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.2553522. This DOI represents all versions, and will always resolve to the latest one. Read more.

doi.org/10.5281/zenodo.2609187

Result



■ Implementation on GitHub

IIIDepence / unarXive

Unwatch 2 Star 0 Fork 0

<> Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master New pull request

Create new file Upload files Find File Clone or download

IIIDepence update Zenodo link Latest commit c615124 8 days ago

doc	add MAG DB schema note	2 months ago
LICENSE	move code to dedicated repo	2 months ago
README.md	update Zenodo link	8 days ago
ald_title.db.placeholder	move code to dedicated repo	2 months ago
clean_txt_output.py	move code to dedicated repo	2 months ago
db_model.py	move code to dedicated repo	2 months ago
extract_contexts.py	move code to dedicated repo	2 months ago
match_bibitems_mag.py	move code to dedicated repo	2 months ago
normalize_arxiv_dump.py	move code to dedicated repo	2 months ago
parse_latex_tralics.py	move code to dedicated repo	2 months ago
prepare.py	move code to dedicated repo	2 months ago
requirements.txt	update Zenodo link	8 days ago

README.md

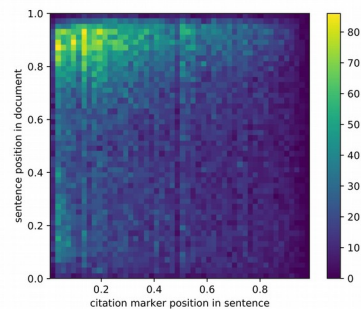
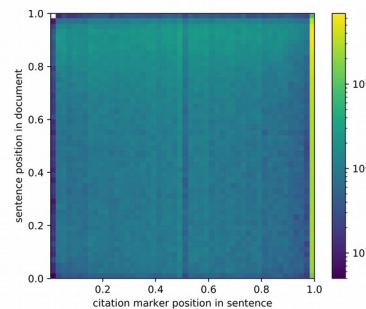
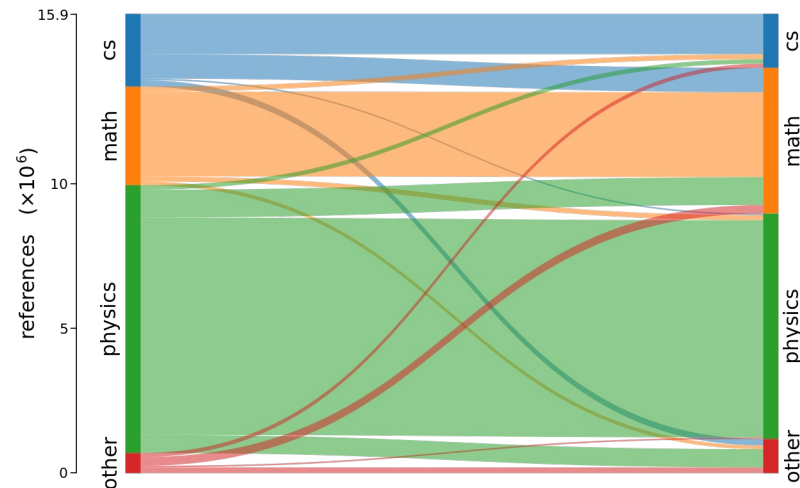
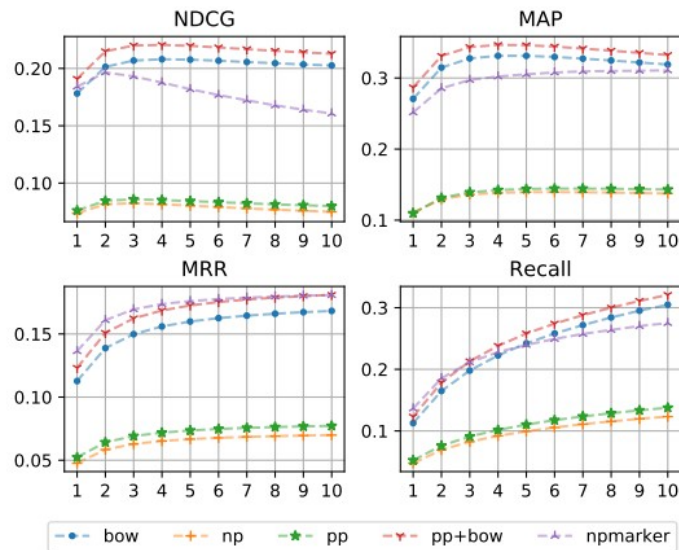
unarXive

Code for generating a data set for citation based tasks using arXiv.org submissions. ([data set on Zenodo](#))

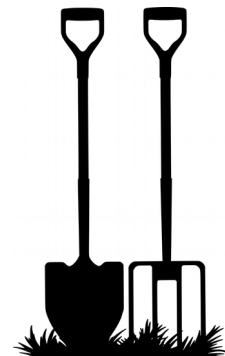
Prerequisites

github.com/IIIDepence/unarXive

Application



Thank you.



References



- **Scholarly** Sugiyama, K., Kan, M.: A Comprehensive Evaluation of Scholarly Paper Recommendation Using Potential Citation Papers. *International Journal on Digital Libraries* 16(2) (2015) 91–109
- **ACL-AAN** Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. *Language Resources and Evaluation* 47(4) (2013) 919–944
- **ACL-ARC** Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M., Lee, D., Powley, B., Radev, D.R., Tan, Y.F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation. LREC'08* (2008)
- **PMC-OAS** <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>
- **PLOS** <https://www.plos.org/text-and-data-mining>
- **CiteSeer^x** Caragea, C., Wu, J., Ciobanu, A.M., Williams, K., Ramírez, J.P.F., Chen, H., Wu, Z., Giles, C.L.: CiteSeer x : A Scholarly Big Dataset. In: *Proceedings of the 36th European Conference on IR Research. ECIR'14* (2014) 311–322
- **arXiv CS** Färber, M., Thiemann, A., Jatowt, A.: A High-Quality Gold Standard for Citation-based Tasks. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation. LREC'18* (2018)
- **MAG** Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: *Proceedings of the 24th International Conference on World Wide Web. WWW'15* (2015) 243–246
- **Tralics** <https://www-sop.inria.fr/marelle/tralics/>
- **Neural ParsCit** Prasad, A., Kaur, M., Kan, M.Y.: Neural ParsCit: A Deep Learning Based Reference String Parser. *International Journal on Digital Libraries* 19 (2018) 323–337