

Pubmed Parser: A Python Parser for PubMed Open-Access XML Subset and MEDLINE XML Dataset XML Dataset

Titipat Achakulvisut¹, Daniel E. Acuna², and Konrad Kording¹

DOI: [10.21105/joss.01979](https://doi.org/10.21105/joss.01979)

1 University of Pennsylvania 2 Syracuse University

Software

- [Review ↗](#)
- [Repository ↗](#)
- [Archive ↗](#)

Editor: Mark A. Jensen [↗](#)

Reviewers:

- [@timClicks](#)
- [@leonardi](#)

Submitted: 12 December 2019

Published: 08 February 2020

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

The number of biomedical publications is increasing exponentially every year. If we had the ability to access, manipulate, and link this information, we could extract knowledge that is perhaps hidden within the figures, text, and citations. In particular, the repositories made available by the [PubMed](#) and [MEDLINE](#) databases enable these kinds of applications at an unprecedented level. Examples of applications that can be built from this dataset range from predicting novel drug-drug interactions, classifying biomedical text data, searching specific oncological profiles, disambiguating author names, or automatically learning a biomedical ontology. Here, we describe Pubmed Parser (`pubmed_parser`), a software to mine Pubmed and MEDLINE efficiently. Pubmed Parser is built on top of Python and can therefore be integrated into a myriad of tools for machine learning such as `scikit-learn` and deep learning such as `tensorflow` and `pytorch`.

Pubmed Parser has the capability of parsing multiple pieces of information into structured datasets that other libraries such as `medic` or `MEDLINEXMLToJSON` do not have. `medic`, for example, does not output paragraphs and captions and has been discontinued since 2015. `MEDLINEXMLToJSON`, similarly, transforms an original XML file into a JSON file, keeping the same structure. It seems also that `MEDLINEXMLToJSON` development has been inactive since 2016. Our parser can be used within Python and provides results in Python dictionaries. It can parse multiple PubMed and MEDLINE data derivatives including article and journal metadata, authors and affiliations, references, figure captions, paragraphs, and more. For example, Pubmed Parser's capabilities were used in Tang et al. (2019) to parse authorship lists, affiliations, and MeSH terms as part of a large-scale name disambiguation pipeline. It has also been used in deep learning pipelines such as one described in Nikolov, Pfeiffer, & Hahnloser (2018), a scientific articles summarization based on titles and abstracts. Parsing XML and HTML with Pubmed Parser allows very efficient production of dictionaries or JSON files that can easily be integrated into downstream pipelines. Moreover, the implemented functions can be scaled easily as part of other MapReduce-like infrastructures such as `PySpark`. This allows users to parse the most recently available corpus and customize the parsing to their needs. Below, we provide an example code to parse an XML file from [MEDLINE](#) corpus.

```
import pubmed_parser as pp
parsed_articles = pp.parse_medline_xml('data/pubmed20n0014.xml.gz',
                                         year_info_only=True,
                                         nlm_category=False,
                                         author_list=False)
```

Pubmed Parser has already been used in published work for several different purposes, including author name disambiguation (Tang et al., 2019), information extraction and summarization

(Abdeddaïm, Vimard, & Soualmia, 2018; Achakulvisut, Bhagavatula, Acuna, & Kording, 2019; Galea, Laponogov, & Veselkov, 2018; Mesbah, Bozzon, Lofi, & Houben, 2018; Nikolov et al., 2018), search engine optimization (Shahri & Kahanda, 2019; Ševa et al., 2019), and biomedical discovery (Miller, 2017; Rakhi, Tuwani, Mukherjee, & Bagler, 2018; Shahri & Kahanda, 2019). It has also been used in multiple biomedical and natural language class projects, and various data science blog posts relating to biomedical text analysis.

Acknowledgements

Titipat Achakulvisut was supported by the Royal Thai Government Scholarship grant #50AC002. Daniel E. Acuna is supported by National Science Foundation grant #1800956.

References

Abdeddaïm, S., Vimard, S., & Soualmia, L. F. (2018). The MeSH-gram Neural Network Model: Extending word embedding vectors with MeSH concepts for UMLS semantic similarity and relatedness in the biomedical domain. *arXiv preprint arXiv:1812.02309*. Retrieved from <https://arxiv.org/abs/1812.02309>

Achakulvisut, T., Bhagavatula, C., Acuna, D., & Kording, K. (2019). Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*. Retrieved from <https://arxiv.org/abs/1907.00962>

Galea, D., Laponogov, I., & Veselkov, K. (2018). Sub-word information in pre-trained biomedical word representations: Evaluation and hyper-parameter optimization. In *Proceedings of the bionlp 2018 workshop* (pp. 56–66). doi:[10.18653/v1/w18-2307](https://doi.org/10.18653/v1/w18-2307)

Mesbah, S., Bozzon, A., Lofi, C., & Houben, G.-J. (2018). SmartPub: A platform for long-tail entity extraction from scientific publications. In *Companion proceedings of the The Web Conference 2018* (pp. 191–194).

Miller, D. (2017). Automated identification of drug-drug interactions in pediatric congestive heart failure patients. *arXiv preprint arXiv:1702.04615*. Retrieved from <https://arxiv.org/abs/1702.04615>

Nikolov, N. I., Pfeiffer, M., & Hahnloser, R. H. (2018). Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*. Retrieved from <https://arxiv.org/abs/1804.08875>

Rakhi, N., Tuwani, R., Mukherjee, J., & Bagler, G. (2018). Data-driven analysis of biomedical literature suggests broad-spectrum benefits of culinary herbs and spices. *PLoS one*, 13(5), e0198030. doi:[10.1371/journal.pone.0198030](https://doi.org/10.1371/journal.pone.0198030)

Shahri, M. P., & Kahanda, I. (2019). ProPheno 1.0: An online dataset for accelerating the complete characterization of the human protein-phenotype landscape in biomedical literature. doi:[10.7287/peerj.preprints.27479v2](https://doi.org/10.7287/peerj.preprints.27479v2)

Ševa, J., Wiegandt, D. L., Götze, J., Lamping, M., Rieke, D., Schäfer, R., Jähnichen, P., et al. (2019). VIST-a variant-information search tool for precision oncology. *BMC bioinformatics*, 20(1), 429. doi:[10.1186/s12859-019-2958-3](https://doi.org/10.1186/s12859-019-2958-3)

Tang, A., Wu, C., Liu, J., Wang, W., Yang, X., & Xing, Y. (2019). Parallel computing for large-scale author name disambiguation in MEDLINE. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 1580–1586). IEEE. doi:[10.1109/hpcc.smartcity.dss.2019.00217](https://doi.org/10.1109/hpcc.smartcity.dss.2019.00217)