The Journal of Open Source Software

LISC: A Python Package for Scientific Literature Collection and Analysis

Thomas Donoghue[1]

1 Department of Cognitive Science, University of California, San Diego

DOI: 10.21105/joss.01674

Software

- Review
- Repository
- Archive

Submitted: 16 August 2019
Published: 26 September 2019

Summary

The scientific literature is vast, and ever expanding. For example, the Pubmed database, a curated database of literature from the bio-medical sciences, holds more than 30 million published scientific articles, and is continuously growing. Given the scale of the literature, work across informatics, information sciences, and bibliometrics has explored automated methods for the curation of and inference from the existing literature. This work is sometimes referred to as knowledge discovery, literature-based discovery, or hypothesis generation (Spangler et al., 2014; Stegmann & Grohmann, 2003; J. B. Voytek & Voytek, 2012).

Here, we introduce 'Literature Scanner', or 'LISC', an open-source Python module for performing automated meta-analyses of scientific articles by collecting and analyzing data from the scientific literature. LISC seeks to provide an easily accessible interface that connects to external resources that make data available through application programming interfaces (APIs). For example, LISC connects to the Pubmed database, providing access to collect and analyze biomedical literature, and to the OpenCitations database (Heibi, Peroni, & Shotton, 2019) providing access to citation data. LISC is designed with an extendable approach that can be used to integrate additional APIs. LISC also includes support and utilities for analyzing the collected literature data.

For data collection, LISC currently offers the following types of literature data collection:

- Counts: tools to collect and analyze data on the co-occurrence of specified search terms
- Words: tools to collect and analyze text and meta-data from scientific articles
- Citations: tools to collect and analyze citation and reference data

To support use cases for collection and analyzing literature data, LISC includes:

- URL management and requesting for interacting with integrated APIs
- custom data objects for managing collected data
- a database structure, as well as save and load utilities for storing collected data
- functions and utilities to analyze collected data
- data visualization for plotting collected data and analysis outputs

LISC is organized as an object-oriented tool, and aims to be a general utility that can be expanded to included new databases, APIs, and analyses as new resources and tools are integrated.

Statement of Need

The size and increasing scale of the scientific literature is prohibitively large for individual scientists to be able to keep up with. Common methods for literature summarization, including meta-analyses and systematic reviews, require time-intensive manual work, and are often limited in scope and lag behind the literature. As a way to complement such approaches, multiple lines of investigation have shown how automated analyses of scientific literature can be applied to summarize and make inference from the existing literature (Spangler et al., 2014; Stegmann & Grohmann, 2003; J. B. Voytek & Voytek, 2012).

Despite these established methods for analyzing the continuously growing literature, there is currently a relative lack of openly available tools to collect and analyze scientific literature. Although databases such as Pubmed have APIs, it still takes considerable work to implement and apply even relatively simple analyses of the literature. LISC seeks to help fill this gap, by providing user-friendly access to methods to programmatically search for and collect literature of interest and apply analyses of interest to it.

LISC aims to serve as a complement to other relevant tools, for example Moliere, a more sophisticated and also more computationally complex tool for hypothesis generation (Sybrandt, Shtutman, & Safro, 2017), or Meta, a recently developed service for probing a pre-built knowledge network inferred from the literature (https://www.meta.org). LISC, in contrast to these more complex systems, aims to offer a lightweight and customizable approach for finding and collecting literature of interest, and offers tools for efficiently performing analyses on this data. It aims to do so in particular by offering a connective interface between available APIs and natural language processing (NLP) analyses available through other tools. The goal is to allow for simple and rapid literature analyses. LISC may not be appropriate for more complex analyses and hypothesis generation projects that would be best served by tools like Moliere.

Related Projects

LISC is inspired by and based on the BRAIN-SCANR project, a project that collected literature data and analyzed co-occurrences of terms in the neuroscientific literature (J. B. Voytek & Voytek, 2012).

LISC and its precursors have enabled a series of recent studies, including meta-analytic / descriptive work and hypothesis driven investigations, such as:

- ERPSCANR: an automated meta-analysis of the field of event-related potential (ERP) work, in the domain of cognitive neuroscience (https://github.com/TomDonoghue/ERP_SCANR).
- Conveyed Confidence in Scientific Literature and Press Releases: an analysis of conveyed confidence in primary scientific literature, as compared to press releases (Fox & Donoghue, 2018)
- Cognitive Ontology Mapping: an analysis of ontologies of cognitive and neuroscientific terms and their use in journal articles and conference proceedings (Gao, Donoghue, & Voytek, 2017)

Acknowledgements

Thank you to Jessica and Bradley Voytek for the inspiration from the BRAINSCANR project, and to Lakshmi Menon, Will Fox and Richard Gao for helpful insights and feedback.

References

Fox, W., & Donoghue, T. (2018). Confidence levels in scientific writing: Automated mining of primary literature and press releases. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1680–1685). Cognitive Science Society.

Gao, R., Donoghue, T., & Voytek, B. (2017). Automated generation of cognitive ontology via web text-mining. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2067–2072). Cognitive Science Society.

Heibi, I., Peroni, S., & Shotton, D. (2019). COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *arXiv*. Retrieved from http://arxiv.org/abs/1904.06052

Spangler, S., Myers, J. N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J. J., Parikh, N., et al. (2014). Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. doi:10.1145/2623330.2623667

Stegmann, J., & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. *Scientometrics*, *56*(1).

Sybrandt, J., Shtutman, M., & Safro, I. (2017). MOLIERE: Automatic biomedical hypothesis generation system. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17 (pp. 1633–1642). ACM. doi:10.1145/3097983.3098057

Voytek, J. B., & Voytek, B. (2012). Automated cognome construction and semi-automated hypothesis generation. *Journal of Neuroscience Methods*, *208*(1). doi:10.1016/j.jneumeth.2012.04.019