Natural Language Processing and the xDeepDive Architecture for Article Recommendation

Simon J Goring Socorro Dominguez Kelly Wu Ty Andrews
Jenit Jain Shaun Hutchinson

Data repositories that rely on users to submit data will often have data holdings that are biased towards individuals who know about the data repository. Inequity in academic resources, and in knowledge networks within academia can mean that data resources that are open, may be magnifying these inequities since not all researchers will have access to the data repository. To help address this inequity we propose a service that can use article metadata to identify articles that may be well suited to the Neotoma Paleoecology Database. This will allow data curators to reach out to individual authors to solicit their contributions, acting as a form of outreach for the database, and helping to build a larger community of contributors to improve the breadth of data holdings within the database.

Table of contents

Introduction

Community Curated Data Resources play an important role in managing and providing data to disciplinary research communities. In particular CCDRs such as Neotoma act as a nexus for education, research, outreach and community by serving as a focal point for the broader community (Goring et al. 2018). The Neotoma Paleoecology Database is a global data resource supporting paleoecological research (Williams et al. 2018). We're building a tool that can scan article text to identify whether or not a paper is suited for inclusion to the Neotoma Paleoecology Database.

Methodology

We're using NLP tools and a machine learning algorithm to identify suitability for a paper within Neotoma. From this we will then extract metadata to create a default object.

Algorithm and Implementation

Article relevance training

We downloaded n papers from PubMed using the pubmed API and a set of keywords that would be likely to include articles relevant to Neotoma, but also articles without direct relevance ("pollen", "archaeology", " 'stone age' ", "aerobiology", "allergies", "mastodon", "diatoms", "paleoecology", "space", "diatom AND paleoecology", "ostracode", "high resolution sediment"). We supplemented this list of articles with all articles within Neotoma that had an accompanying DOI. All articles were then hand-tagged (by SJG) using SMART () as either "Neotoma", "Not Neotoma" and "Maybe Neotoma". We used the "Maybe Neotoma" tag for articles that were likely to be of interest to the Neotoma Data Steward community, but were unlikely to be entered into Neotoma because the primary disciplinary community likely had a different data repository of record. For example, high resolution tephra-chronology is critical for chronology construction within Neotoma, but the primary repository of record is likely EarthChem.

From the tagged articles we extracted metadata from CrossRef and PubMed to provide a more complete data object. This metadata excluded the use of the article fulltext since this would not be available for many legacy publications. Using an XXXX model on a set of publisher

supplied features we then constructed the model to both predict article suitability, and also the obtain feature importance rankings for the article.

Results

Article Tagging

The dataset is highly imbalanced. Of the n total articles tagged, only n_2 articles were deemed suited for Neotoma, with n_3 identified as being "Maybe Neotoma" and n_4 being "Not Neotoma". This reflects the fact that paleoecology papers that include primary data represent only a small proportion of all articles published.

Article Relevance

The final model was $x\%$ accurate, identifying stuff? Feature importance . . .

Conclusions

Acknowledgements

Code availability Section

- Contact: goring@wisc.edu
- Hardware requirements: IBM PC 8086 Processor with 1MB RAM and two 1.44MB Floppy Disks.
- Program language: Python
- Software required: …
- Program size: …

The source codes are available for downloading at the link: https://github.com/NeotomaDB/MetaExtractor

References

Goring, Simon James, Russell Graham, Shane Loeffler, Amy Myrbo, James S. Oliver, Carol Ormond, and John W. Williams. 2018. *The Neotoma Paleoecology Database: A Research Outreach Nexus*. Elements of Paleontology. Cambridge University Press. https://doi.org/10.1017/9781108681582.

Williams, John W, Eric C Grimm, Jessica L Blois, Donald F Charles, Edward B Davis, Simon J Goring, Russell W Graham, et al. 2018. "The Neotoma Paleoecology Database, a Multiproxy, International, Community-Curated Data Resource." *Quaternary Research* 89 (1): 156–77. https://doi.org/10.1017/qua.2017.105.