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

```
Warning in readLines(file): incomplete final line found on '.env'

Warning in readLines(file): incomplete final line found on '.env'
```

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). Data representivity can be a considerable challenge for CCDRs as a result of bias in data contributors, or the practices and global distributions of researchers and research projects (Zhao et al. 2019; Queenan et al. 2016; Proença et al. 2017; Minor et al. 2016). Biases in data representivity can be addressed in several way. Organizations such as the International Tree Ring Database have used statistical tools to identify high priority areas for future study (Zhao et al. 2019). This approach provides high level guidance to existing users of the data resource, providing them with an authority to use in requesting research funding and planning future data collection campaigns. Medical databases can use population-level resources such as census data to identify shortcomings within their data resources (Queenan et al. 2016), prompting calls for additional recruitment of data contributors with patient-bases that are more representative of the overall population. In general these two approaches lead us to either targeting existing data (or ongoing data collection efforts) or targeting new data collection efforts through advocay.

The Neotoma Paleoecology Database is a global data resource supporting paleoecological research (Williams et al. 2018). Neotoma is a database of databases, representing many distinct user communities and data collection efforts, sharing a common data model and leadership structure. Neotoma represents data contributions from XXXX researchers from around the globe, however, Neotoma relies entirely on user contributions, and as such is susceptible to data representivity bias.

One approach to addressing representation bias is to proactively request data contributions from researchers when articles that are well suited for inclusion in the database appear in the

literature. A challenge with this approach is that it requires one or more individuals to be attentive to article alerts across a number of potential search terms. Neotoma contains data from at least 25 different data-types, including pollen, diatoms, water chemistry and vertebrate fossils, however, not all papers published on these subjects is suited to Neotoma. This means that there would be considerable effort required to search, filter and request data from the appropriate papers.

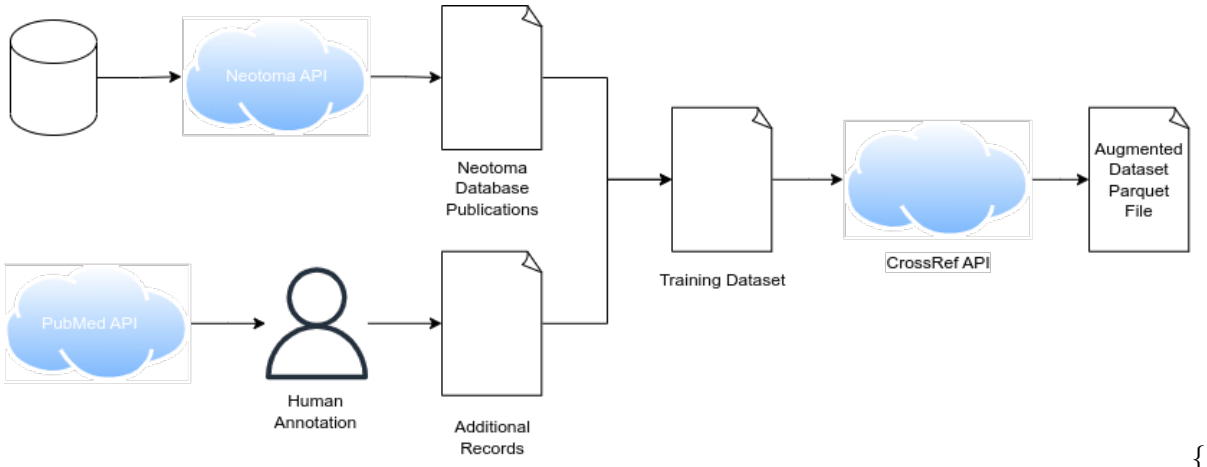
Advances in machine learning and new sources of digital data point to a way forward using a data science approach. Leveraging full text searching and rich metadata it should be possible to provide an inclusive set of search terms and allow a machine learning algorithm to predict article relevance. Here we present such an approach. Using the GeoDeepDive server we search for all articles published within a certain time period with a set of known terms. We augment the article metadata with additional metadata from CrossRef to provide a set of data on which a predictive model can be applioed. A predictive model is then applied to the set of article metadata to indicate relevance along a 0 - 1 score. These articles can then be easily assessed and evaluated for inclusion within the 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. The approach described here uses human curation, and data from the Neotoma Paleoecology Database to build a set of publication metadata that can be used to train a Machine Learning algorithm to predict article suitability for the database. Using commercial cloud computing services and public APIs we query data to add to the list of suitable and non-suitable articles for the database, and, with additional manual curation, we re-train the model to (ideally) improve model outputs in the long-term.

Algorithm and Implementation

Article relevance training



{100%}

```
Warning: One or more parsing issues, call `problems()` on your data frame for details,
e.g.:
  dat <- vroom(...)
  problems(dat)
```

We downloaded 1844 papers from PubMed using the PubMed API ((**PubmedAPI?**)) 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"). The list of PubMed sourced articles was supplemented by a list of all articles within Neotoma that had an accompanying DOI 579. All articles were then hand-tagged (by SJG) using SMART (Chew et al. 2019) 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 tephrachronology 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. For each model we apply an 80/20 training-testing split to the data to limit the possibility of overfitting. A set of baseline models were then constructed using SciKitLearn in Python.

All articles, with associated metadata were then stored in a Parquet formatted file in an AWS S3 bucket in the cloud. Each article included all associated metadata and the additional columns "relevant", "...".

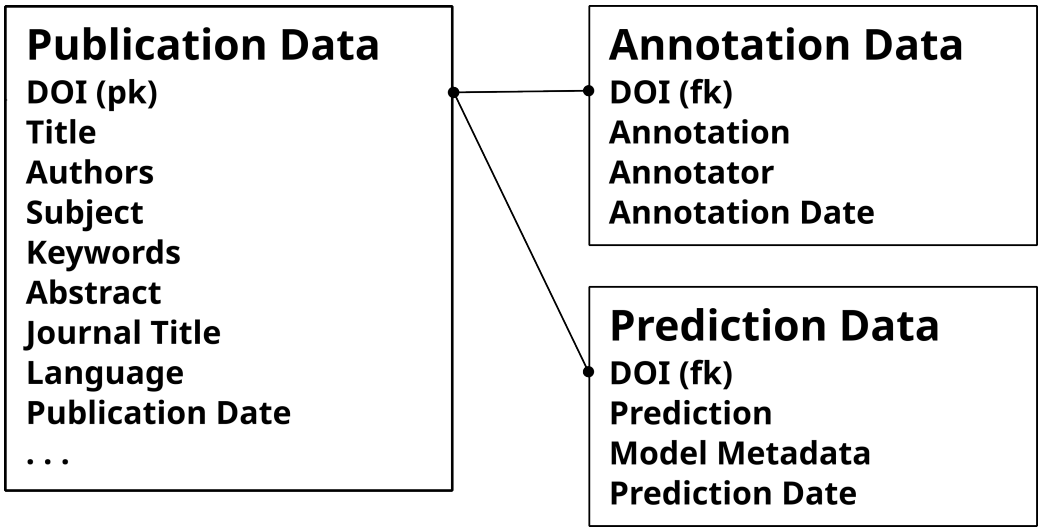


Figure 1: Data structure for this project includes a single file containing all article metadata, drawn from CrossRef and Web of Science APIs. Each article may be annotated by an individual or have known provenance from the Neotoma Paleoecolgy Database using the DOI as a foreign key to tie it to the article metadata. All predictions are associated with the DOI, and are presented in an independent file to allow us to also test how changes in the underlying model may result in changes in overal prediction ability.

Model building used nine models implemented from the Python ScikitLearn package (Pedregosa et al. 2011), along with a dummy model against which to compare models. For each model we assessed the time required to fit the model and to predict results on the testing dataset, recall for the test and training sets, F1 scores for the test and training sets, as well as precision and accuracy for the datasets.

The models are trained on the article subject (as defined by the publisher and reported by CrossRef), and a "bag of words" representation of the concatenated article title and abstract. We do not report "journal" because we want to be able to report the relevance of articles in new (or old) journals that are not represented in the training dataset, particularly if these journals represent regional or domain specific journals that represent underserved communities. We use "Subject" as a proxy for journal focus or content. We concatenate article title and abstract to capture greater textual information about the articles. Additionally, metadata reporting by publishers to CrossRef is variable. We almost always get an article title, but, for particular journals, abstract is often excluded.

Data Preprocessing

Both the Subject column and the Title/Abstract concatenation are pre-processed using `CountVectorizer()` to generate a sparse matrix of terms associated with each entry. We limit the maximum number of terms to 1000, remove english stopwords and remove accents using unicode mapping. In addition, the default parameters for `CountVectorzer()` transform all terms to lowercase.

Model Selection and Tuning

From the initial model training exercise

Results

Article Tagging

Training Data

Neotoma maintains an internal list of publications associated with the database datasets. The 47796 datasets within Neotoma are associated with 8431 publications, spanning the last y years. Of the records within Neotoma, many come from grey literature, or were added to the database before the wide use of DOIs for articles. As a result there is an element of incomplete data across the records. For this task, we limited the articles to only those with a recorded DOI in the database. This resulted in the inclusion of 1237 articles from Neotoma. From there, an additional 1237 articles were obtained and tagged to improve the model. These were tagged using the SMART (Chew et al. 2019) annotation tool, with classes "Neotoma", "Not Neotoma" and "Maybe Neotoma".

The composite dataset is highly imbalanced. Of the 2423 articles tagged (or already within Neotoma), only 671 articles were classed as being appropriate for Neotoma, with 73 identified as being "Maybe Neotoma" and 1679 were in the class "Not Neotoma". This reflects the fact that paleoecology papers including primary data represent only a small proportion of all articles published, even when subject and keyword searches focus on terms suited to the discipline.

```
`summarise()` has grouped output by 'doi', 'journal', 'label'. You can override
using the `.groups` argument.
`summarise()` has grouped output by 'journal', 'label'. You can override using
the `.groups` argument.
`summarise()` has grouped output by 'doi', 'subject', 'label'. You can override
using the `.groups` argument.
```

```
`summarise()` has grouped output by 'subject', 'label'. You can override using
the `.groups` argument.
```

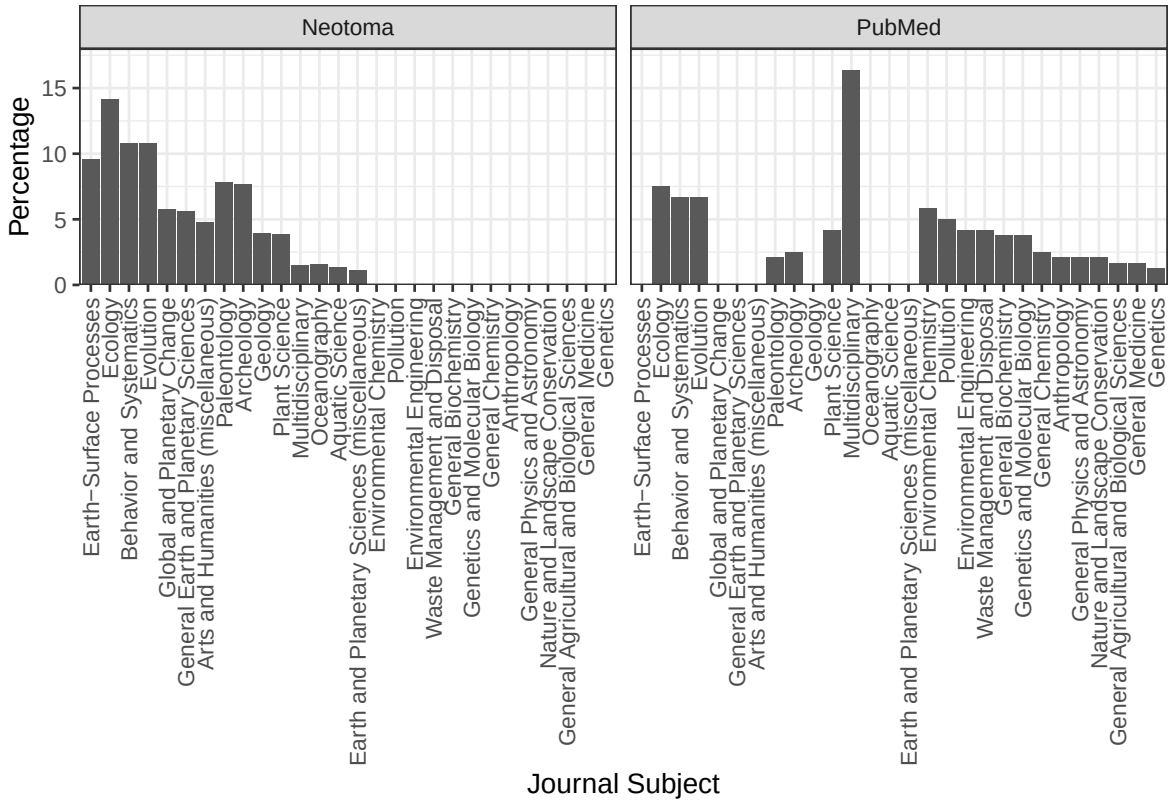


Figure 2: Distribution of journal subjects identified from journals with articles contained in Neotoma and articles from the training set obtained from PubMed, expressed as percentages. The PubMed records show increased representation by subjects that are not traditionally 'ecology' based, including 'pollution', 'anthropology' and 'environmental chemistry'. These represent subjects associated with emerging constituent groups within Neotoma, and may indicate new sources and communities of data for Neotoma.

Training data with broader coverage of subject matter and journals will be likely to capture a broader range of abstract construction and key terms, providing us with a more robust dataset for training. Within the existing papers we see evidence of skew in journal and subject representation within Neotoma and within the training set. In part this is a result of bias in the PubMed journal holdings (not all Earth Science journals are indexed within PubMed), but it is also a result of skew in the source journals submitted to Neotoma, and the community that Neotoma draws from. Quaternary Research, The Holocene and Quaternary Science Reviews represent the largest proportion for Neotoma records, but these primarily represent the tradi-

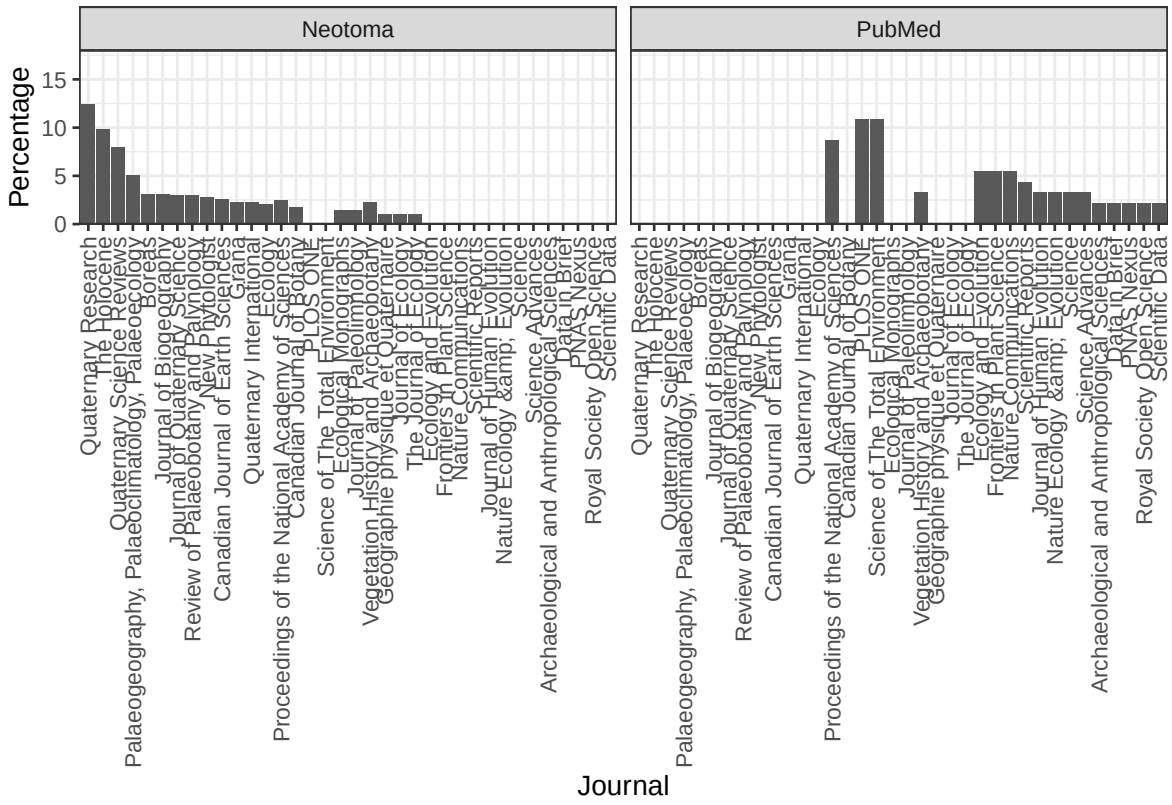


Figure 3: Distribution of journal titles identified from articles contained in Neotoma and articles from the training set obtained from PubMed, expressed as percentages. The PubMed records show increased representation by multidisciplinary journals (PNAS, PLOS ONE, Nature Communications) as opposed to the articles sources from Neotoma.

tional paleoecology community. Emerging research communities may be publishing in either multi-disciplinary journals (PLOS ONE, Scientific Reports), or may be publishing data papers directly (Data in Brief, Scientific Data).

Model Building

Model building used nine models implemented from the Python ScikitLearn package (Pedregosa et al. 2011), along with a dummy model against which to compare models. For each model we assessed the time required to fit the model and to predict results on the testing dataset, recall for the test and training sets, F1 scores for the test and training sets, as well as precision and accuracy for the datasets.

Table X. *Models used for binary ('relevant', 'not relevant') relevance fitting. Each model was run with a* `random_state` *defined, to ensure reproducibility of results, but the state is not defined in this table, to save space.*

Model	Model Class	Function Call
Dummy	Null	DummyClassifier()
Logistic Regression	Linear Model	LogisticRegression(class_weight="balanced", max_iter=1000)
Decision Tree	Decision Tree	DecisionTreeClassifier(class_weight="balanced", max_depth=200)
kNN	Nearest Neighbours	KNeighborsClassifier()
Naive Bayes	Naive Bayes	BernoulliNB()
RBF SVM	Support Vector Machine	SVC(class_weight="balanced")
RF	Ensemble Methods	RandomForestClassifier(class_weight="balanced")
LGBM	Gradient Boosting	LGBMClassifier(class_weight="balanced")
CatBoost	Gradient Boosting	CatBoostClassifier(verbose=0)
XGBoost	Gradient Boosting	XGBClassifier(class_weight="balanced", verbosity=0)

Article Relevance

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

Subject Representivity

TODO: Test representivity by comparing the distribution of journal articles by journal; between recommended and the current distribution in Neotoma (?)

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 (MetaReview package), R (this paper)
- Software required: None
- Program size: …

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

References

Chew, Rob, Michael Wenger, Caroline Kery, Jason Nance, Keith Richards, Emily Hadley, and Peter Baumgartner. 2019. "SMART: An Open Source Data Labeling Platform for Supervised Learning." *Journal of Machine Learning Research* 20 (82): 1–5. http://jmlr.org/papers/v20/18-859.html.

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.

Minor, Wladek, Zbigniew Dauter, John R Helliwell, Mariusz Jaskolski, and Alexander Wlodawer. 2016. "Safeguarding Structural Data Repositories Against Bad Apples." *Structure* 24 (2): 216–20.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Proença, Vânia, Laura Jane Martin, Henrique Miguel Pereira, Miguel Fernandez, Louise McRae, Jayne Belnap, Monika Böhm, et al. 2017. "Global Biodiversity Monitoring: From Data Sources to Essential Biodiversity Variables." *Biological Conservation* 213: 256–63. https://doi.org/https://doi.org/10.1016/j.biocon.2016.07.014.

Queenan, John A, Tyler Williamson, Shahriar Khan, Neil Drummond, Stephanie Garies, Rachael Morkem, and Richard Birtwhistle. 2016. "Representativeness of Patients and Providers in the Canadian Primary Care Sentinel Surveillance Network: A Cross-Sectional Study." *Canadian Medical Association Open Access Journal* 4 (1): E28–32.

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.

Zhao, Shoudong, Neil Pederson, Loïc D'Orangeville, Janneke HilleRisLambers, Emery Boose, Caterina Penone, Bruce Bauer, Yuan Jiang, and Rubén D Manzanedo. 2019. "The International Tree-Ring Data Bank (ITRDB) Revisited: Data Availability and Global Ecological Representativity." *Journal of Biogeography* 46 (2): 355–68.