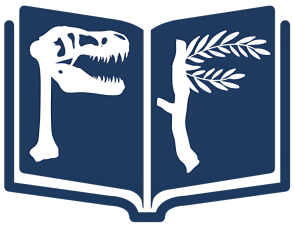
Proposal:
Finding Fossils in the Literature



In partnership with the Neotoma Paleoecology Database.

Ty Andrews Jenit Jain Kelly Wu Shaun Hutchinson

5/12/23

Executive Summary

Finding Fossils in the Literature is sponsored by the Neotoma database (Neotoma) which houses paleoecology data (e.g. excavation site locations, taxa, etc.). The challenges Neotoma faces are 1) researchers have to manually enter sample data into Neotoma, 2) researchers are not aware of Neotoma or that their research fits into it, and 3) there are too many articles published for the Neotoma team to monitor new research. This project has 3 primary deliverables to solve the challenges, first is an article relevance prediction model which predicts whether newly published articles are relevant to Neotoma. Second, is an article data extraction pipeline which identifies key entities such as taxa or geographic location. Last is a data review tool for Neotoma data stewards to review the extracted data before it is submitted to Neotoma. A mid-project demo is set as a goal for the week of May 29 to have a functioning MVP for each data product with final versions completed by June 15 and project documentation and handoff completed by June 28.

Table of contents

1 Introduction

The Neotoma database (Neotoma) (Williams et al. 2018) is used by researchers studying ecological changes over the past 5 million years. However, the data collection process relies heavily on manual submissions by researchers, leading to challenges in data entry and hindering collaborative efforts to comprehend ecological changes comprehensively. This project aims to automate the extraction of data from relevant journal articles which can be added to Neotoma. This will be done in three parts. First article relevancy to the Neotoma will be predicted. Relevant articles will be parsed using natural language processing (NLP) techniques. Finally a tool will be built to review the extracted data before it is submitted to Neotoma.

2 Article Relevance Prediction

The first step is to build a document classification model to assess the relevance of the new articles to Neotoma.

2.1 Data

The data to be used for developing the article relevance prediction model comes from the public CrossRef application programming interface (API) which provides data for published journal articles. Articles already in Neotoma are used as the positive examples and non-relevant keyword queries against the CrossRef API will be used for extracting negative examples. Currently, there are 758 articles from Neotoma and the team plan to collect more from both Neotoma and Crossref API to build a representative balanced sample. The data will be preprocessed as follows in Table 1:

Table 1: Proposed preprocessing for article data from the CrossRef API

Variable	**Description**	**Preprocessing**
abstract	Abstract of the article	Text count vectorized
author	Author of the article	One-hot encoding, concatenate author's first initial/last name
container-title	Title of the article's container	One-hot encoding
is-referenced-by-count	Number of references by other articles	Standard scaling
language	Article language	One-hot encoding
published	Date article was published	Year as numeric features |
publisher	Publisher name	One-hot encoding
subject	Subject of the article	Text count vectorized
subtitle	Subtitle of the article	Text count vectorized
title	Title of the article	Text count vectorized

2.2 Proposed Approaches & Success Criteria

It is proposed that supervised machine learning approaches are used to predict article relevancy.

2.2.1 Success Criteria

Approaches will be evaluated primarily on F1-Score with recall being monitored to reduce false negatives (Table 2).

Table 2: Proposed evaluation metric and target value for article relevancy prediction.

Metric	Target
F1-Score	$> 83\%$ (Alex et al. 2022)

2.2.2 Baseline

The baseline approach for this model will be to use logistic regression with a bag of words representation of the article features extracted from the CrossRef API.

2.2.3 Approach 1: Traditional Machine Learning Models

Exisiting research (Tran Thanh, Loc, and Thai-Nghe 2019; also Weber et al. 2020) has shown the following models to perform well on this type of text based classification problem:

- Naive Bayes
- SVM
- Random Forest/Gradient Boosting

The above models are proposed as they can represent non-linear relationships from the text and can manage highly sparse data.

2.2.4 Approach 2: Transfer Learning with BERT Models

Additionally, we will leverage pre-trained BERT based large language models for text embeddings for feature engineering. We will explore transfer learning to fine-tune the BERT pre-trained model so that it better learns the contextual information represented in paleoecology-related articles.

3 Fossil Data Extraction Pipeline

The fossil data extraction pipeline receives the list of articles which are predicted to be relevant and processes each article's full text to pull out data that fits within the Neotoma DB tables.

3.1 Data

The data for the fossil data extraction comes from GeoDeepDive and contains all the text from each article and is received as a list of sentences from GeoDeepDive. To generate labelled entities, we propose using a privately hosted version of LabelStudio(*Label Studio: Data Labeling Software* (version 1.7.3) 2023) on the HuggingFace hub (*HuggingFace* (version 4.29.1) 2023) with labeling done by team members.

The entities to be extracted and their general formats are shown in Table 3:

Table 3: Proposed entities to be labelled in the articles.

Entity Name	**Description**
Geographic Location - GEOG	Longitude/longitude coordinates
Site Name - SITE	Name of the excavation site
Taxa - TAXA	Taxa of samples collected
Age - AGE	Dated age of the samples
Altitude - ALTI	Altitude where sample was collected
Email Address(es) - EMAIL	Email addresses of the researchers

3.2 Proposed Approaches & Success Criteria

It is proposed that named entity recognition (NER) approaches are used to extract the data from the articles.

3.2.1 Success Criteria

Approaches will be evaluated primarily on F1-Score with recall being monitored to reduce false negatives (Table 4).

Table 4: Proposed evaluation metrics and target value for fossil data extraction.

Metric	Target
Micro F1-Score	$> 85\%$ (Conneau et al. 2020)

3.2.2 Baselines

For each entity to be extracted the following approaches are proposed in Table 5.

Table 5: Proposed baseline approach for each entity.

Entity Name	**Baseline Approach**
Geographic Location - GEOG	Regular Expressions (Goring et al. 2021)
Site Name - SITE	spaCy Pre-Trained NER model identifying location entities (*spaCy NER* (version 3.5) 2023)
Taxa - TAXA	In-text search for existing taxa already in Neotoma
Age - AGE	Regular Expressions (Goring et al. 2021)
Altitude - ALTI	Regular Expressions ("above sea level", "a.s.l.")
Email Address(es) - EMAIL	Regular Expressions

3.2.3 Approach 1: Fine Tuned SpaCy NER Model

The spaCy Python package (*spaCy NER* (version 3.5) (2023)) includes the en_core_web_lg NER model. This model utilizes convolutional neural networks to generate text embeddings, which are used to classify each token of text according to the custom entity labels specific to the dataset. It has been pre-trained on texts from the English language and achieves NER accuracy of 85.5% on the OntoNotes 5.0 corpus (Ralph Weischedel 2013). The spaCy Python package (*spaCy NER* (version 3.5) (2023)) includes the en_core_web_lg NER model. This model utilizes convolutional neural networks to generate text embeddings, which are used to classify each token of text according to the custom entity labels specific to the dataset. It has been pre-trained on texts from the English language and achieves NER accuracy of 85.5% on the OntoNotes 5.0 corpus (Ralph Weischedel 2013).

3.2.4 Approach 2: Fine Tuned Transformer NER Model

Two transformer based approaches are proposed to be evaluated. The first is the Text-To-Text Transfer Transformer (T5) model which is unique in that it is a prompt based model

that accepts text then generates text at the output. The second model is the XLM-RoBERTa (XLM-R) model which is a cross-language BERT based model which has the advantage that it is able to handle multi-language inputs which is desirable as some example papers are published in French and other languages.

4 Data Review Tool

The final step in the process is data stewards from Neotoma reviewing the extracted data. It is proposed to build a data review tool using Plotly Dash. It is expected that use of the dashboard should not require any technical software development experience and ongoing maintenance will be managed by Simon Goring and his team.

4.1 Success Criteria

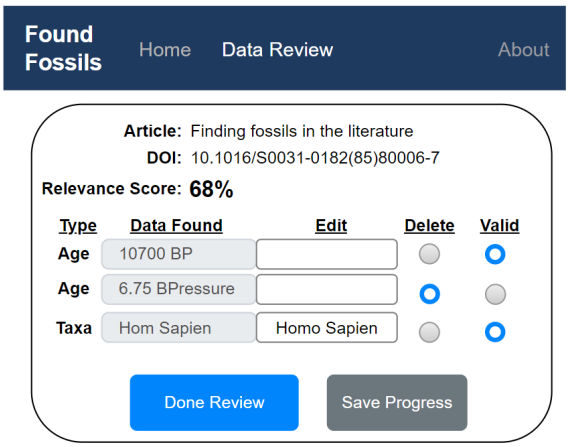
Requirements are summarized in Table 6.

Table 6: Proposed requirements for the data review tool.

Requirement	**Target**
Options for reviewing extracted data	Accept, Reject, Edit then Accept
Other data made available to the user	Article DOI, Hyperlink to Article
Extracted data context provided	Current sentence and 1-2 sentences before/after
User skill to use	Non-Technical (e.g. no code/CLI)
Number of mouse clicks to review single piece of data	1-2
Reviewing workflow	Able to save/resume progress
Output file format	JSON

A draft wireframe for how the tool may look is below in Figure 1.

Figure 1: Data review tool wireframe.

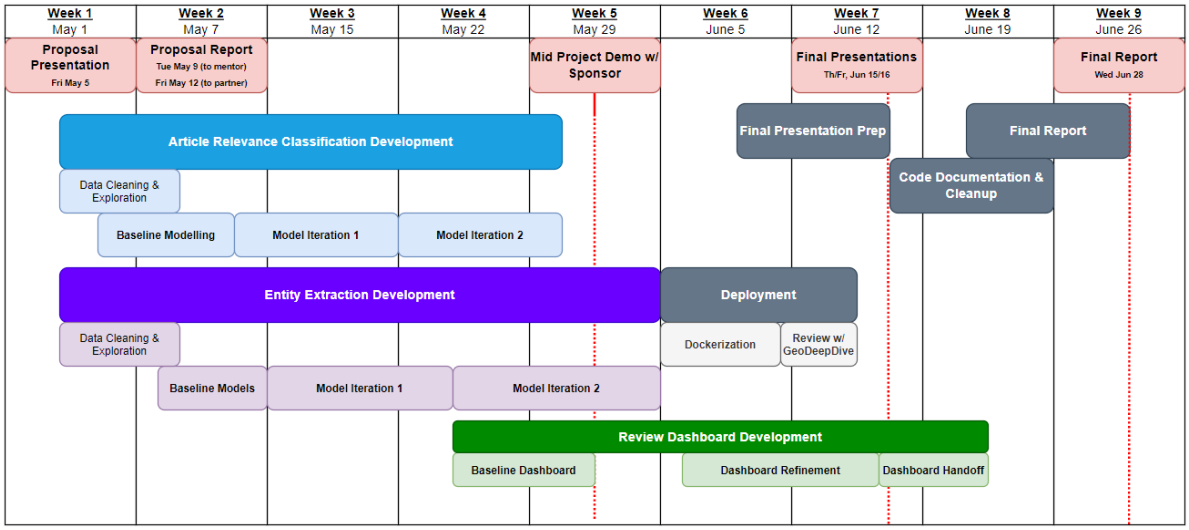


5 Timeline

Deadlines and proposed intermediate milestones are outlined below and in Figure 2. Tasks will be completed in parallel by the team where appropriate.

- *Milestone 1 - May 12th*: Initial data cleaning and baselines complete.
- *Milestone 2 - May 19th*: First iterations of each model complete.
- *Milestone 3 - May 26th*: Second model iterations complete and MVP data review tool built.
- *Mid-project demo*: (Tentative) Show end to end demo and get feedback.
- *Milestone 4 - June 9th*: Solution deployment & final presentation.

Figure 2: Proposed project timeline



6 Acknowledgements

Data were obtained from the Neotoma Paleoecology Database (http://www.neotomadb.org) and its constituent database(s). The work of data contributors, data stewards, and the Neotoma community is gratefully acknowledged.

A huge thanks to Simon Goring & Socorro Dominguez from the Neotoma Database team for their support on this project thus far.

References

Alex, Neel, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, et al. 2022. "RAFT: A Real-World Few-Shot Text Classification Benchmark." https://arxiv.org/abs/2109.14076.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." https://arxiv.org/abs/1911.02116.

Goring, Simon, Jeremiah Marsicek, Shan Ye, John Williams, Stephen Meyers, Shanan Peters, Daven Quinn, Allen Schaen, Brad Singer, and Shaun Marcott. 2021. "A Model Workflow for GeoDeepDive: Locating Pliocene and Pleistocene Ice-Rafted Debris," July. https://doi.org/10.31223/X54312.

HuggingFace (version 4.29.1). 2023. https://huggingface.co/.

Label Studio: Data Labeling Software (version 1.7.3). 2023. https://github.com/heartexlabs/label-studio.

Ralph Weischedel, Mitchell Marcus, Martha Palmer. 2013. "OntoNotes Release 5.0." https://doi.org/https://doi.org/10.35111/xmhb-2b84.

spaCy NER (version 3.5). 2023. Explosion. https://spacy.io/api/entityrecognizer.

Tran Thanh, Dien, Bui Loc, and Nguyen Thai-Nghe. 2019. "Article Classification Using Natural Language Processing and Machine Learning," November, 78–84. https://doi.org/10.1109/ACOMP.2019.00019.

Weber, Tobias, Dieter Kranzlmüller, Michael Fromm, and Nelson Tavares de Sousa. 2020. "Using supervised learning to classify metadata of research data by field of study." *Quantitative Science Studies* 1 (2): 525–50. https://doi.org/10.1162/qss_a_00049.

Williams, J. W., E. G. Grimm, J. Blois, D. F. Charles, E. Davis, S. J. Goring, R. Graham, et al. 2018. "The Neotoma Paleoecology Database: A Multi-Proxy, International Community-Curated Data Resource." *Quaternary Research* 89: 156–77.