TEXT MINING SCHOLARLY PUBLICATIONS USING APIS

Ishita Sarraf, Grinnell College ’25

Information Quality Lab
School of Information Sciences
University of Illinois Urbana-Champaign

METSTI 2023: Workshop on Informetric, Scientometric, and Scientific and Technical Information Research

October 27, 2023

ILLINOIS

RESEARCH QUESTION

How can I create an automated pipeline that will make it easy to deal with copyright licenses so that researchers can create custom datasets easily?

THE PROBLEM

Researchers want to create custom datasets to mine and analyze publications (Bertin & Atanassova, 2018).

They need access to full text of digital publications that have copyright licenses.

Dealing with copyright licenses is time-consuming and difficult.

THE PROBLEM

Researchers want to create custom datasets to mine and analyze publications (Bertin & Atanassova, 2018).

They need access to full text of digital publications that have copyright licenses.

Dealing with copyright licenses is time-consuming and difficult.

THE PROBLEM

Researchers want to create custom datasets to mine and analyze publications (Bertin & Atanassova, 2018).

They need access to full text of digital publications that have copyright licenses.

Dealing with copyright licenses is time-consuming and difficult.

METHODS

DATA COLLECTION PIPELINE

01

Get the DOI from the user and supply it to the Crossref TDM API.

DATA COLLECTION PIPELINE

01

Get the DOI from the user and supply it to the Crossref TDM API.

02

Get the license URL and full text URLs.

DATA COLLECTION PIPELINE

01

Get the DOI from the user and supply it to the Crossref TDM API.

02

Get the license URL and full text URLs.

03

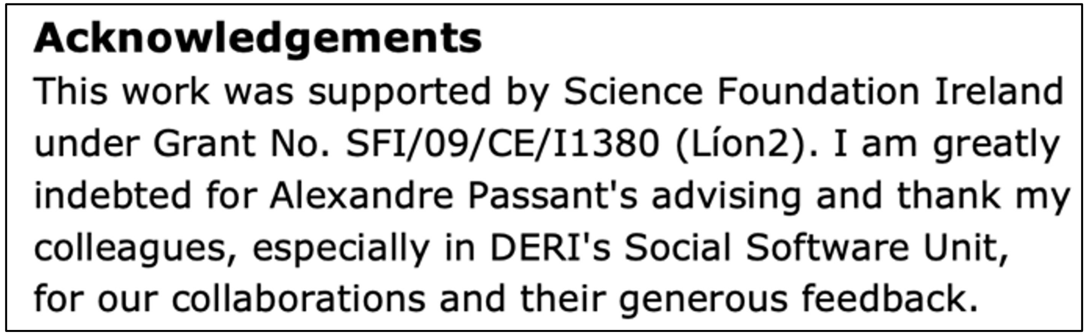
Store the full text (if available) in the database.

REQUIREMENTS ANALYSIS

To identify tasks that can use my pipeline, I am interviewing researchers who mine and analyze publications.

RESULTS

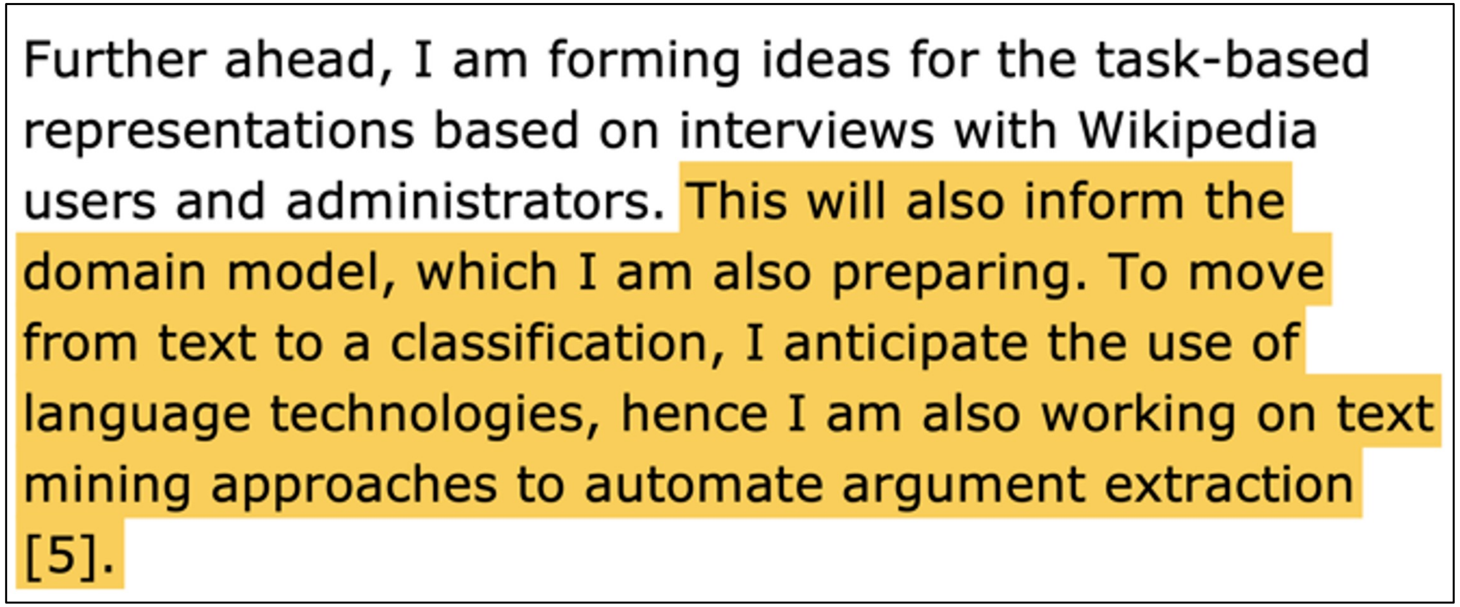
TASK 1: FUNDER INFORMATION EXTRACTION



Acknowledgements

This work was supported by Science Foundation Ireland under Grant No. SFI/09/CE/I1380 (Líon2). I am greatly indebted for Alexandre Passant's advising and thank my colleagues, especially in DERI's Social Software Unit, for our collaborations and their generous feedback.

The Acknowledgments section will be extracted from papers to detect funding bias.



Further ahead, I am forming ideas for the task-based representations based on interviews with Wikipedia users and administrators. This will also inform the domain model, which I am also preparing. To move from text to a classification, I anticipate the use of language technologies, hence I am also working on text mining approaches to automate argument extraction [5].

TASK 2: CITATION CONTEXT ANALYSIS

- The highlighted text shows sentences around an in-text citation.
- These sentences will be used for citation context analysis.

CODE GLITCH IN A COMPUTATIONAL CHEMISTRY PROTOCOL PROJECT

287 DOIs were tested on the pipeline to download full text

30% DOIs were downloaded with full text

FUTURE WORK AND CONCLUSION

I will continue to develop the pipeline to implement these applications.

I plan to work on fixing the errors found in the Computational Chemistry Protocol Project

With further requirements analysis, I will investigate and incorporate additional possible applications for analyzing full texts of various scholarly publications.

ACKNOWLEDGMENTS

Ishita Sarraf was supported by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA-W) and the Coalition to Diversify Computing (CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF BPC-A #1246649).

My mentors: Dr. Jodi Schneider and Yuanxi Fu are funded by NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis.

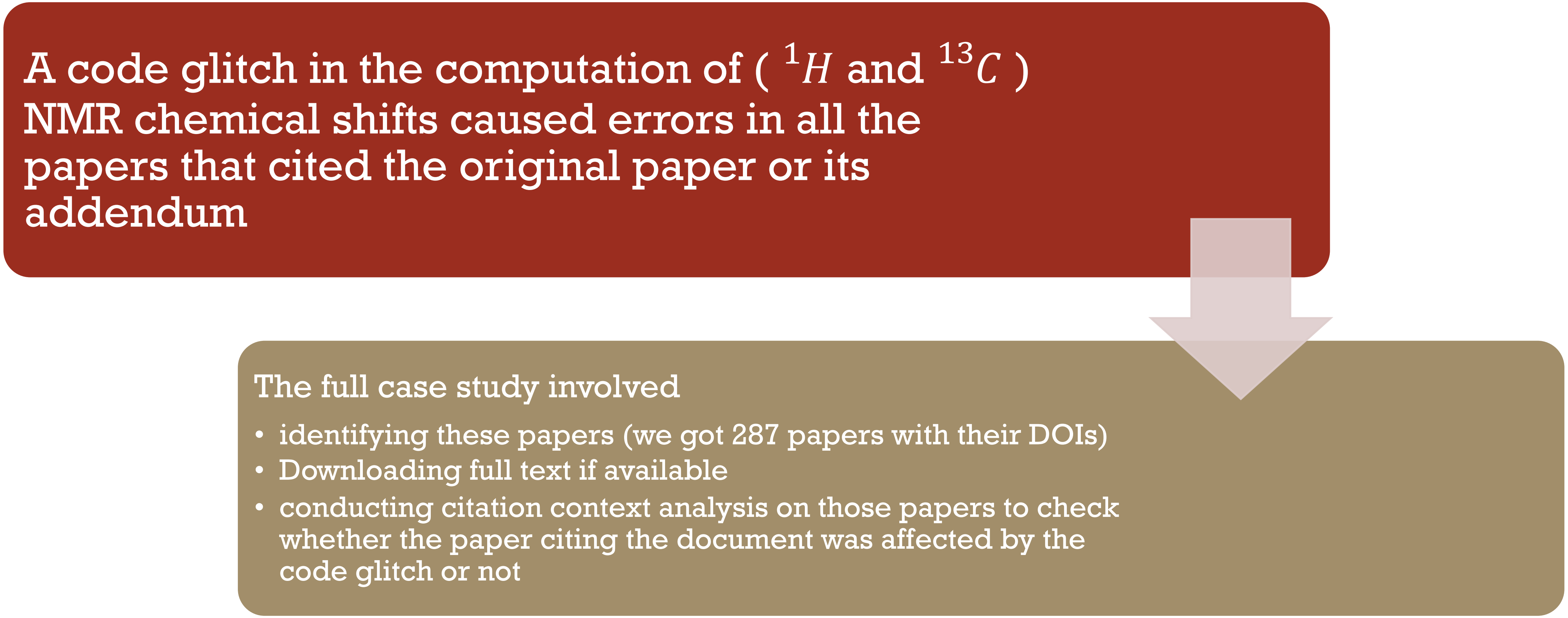
My family and friends.

REFERENCES

- American Psychological Association. (2020, August). What is a Digital Object Identifier, or DOI? https://apastyle.apa.org/learn/faqs/what-is-doi
- Berners-Lee, T. (2009, August 27). Metadata Architecture. https://www.w3.org/DesignIssues/Metadata.html
- Bertin, M., & Atanassova, I. (2018). InTeReC: In-text Reference Corpus for applying natural language processing to bibliometrics. In P. Mayr, I. Frommholz, & G. Cabanac (Eds.), *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval.* (Vol. 2080, pp. 54–62). CEUR. https://ceur-ws.org/Vol-2080/paper6.pdf
- Biehl, M. (2015). API Architecture: The Big Picture for Building APIs. https://restapilinks.com/wp-content/uploads/2021/02/api_architecture_biehl.pdf

- Himmelfarb Health Sciences Library. (2023, May 24). Scholarly Publishing: Scholarly Publishing. https://guides.himmelfarb.gwu.edu/scholarlypub
- Lammey R. (2014). CrossRef's text and data mining services. *Learned Publishing*, 27(4), 245–250. https://doi.org/10.1087/20140402
- Polischuk, P. (2020, April 8). Text and data mining. Crossref. Retrieved July 5, 2023 from https://www.crossref.org/documentation/retrieve-metadata/rest-api/text-and-data-mining/
- Vickery, B. (2021, August 21). Evolving our support for text-and-data mining. Crossref. Retrieved July 5, 2023 from https://www.crossref.org/blog/evolving-our-support-for-text-and-data-mining/

CODE GLITCH IN A COMPUTATIONAL CHEMISTRY PROTOCOL PROJECT



A code glitch in the computation of (^{1}H and ^{13}C) NMR chemical shifts caused errors in all the papers that cited the original paper or its addendum

The full case study involved

- identifying these papers (we got 287 papers with their DOIs)
- Downloading full text if available
- conducting citation context analysis on those papers to check whether the paper citing the document was affected by the code glitch or not

REASONS FOR NOT GETTING FULL TEXT

1

INSTITUTION DOES NOT HAVE ACCESS

2

XML or PDF URLs NOT AVAILABLE/ UNSPECIFIED URLS

3

PUBLISHER NOT GIVING PERMISSION

ILLINOIS