Ishita Sarraf
August 15, 2023
Final Report
Distributed Experience for Undergraduates Program
Information Quality Lab, University of Illinois at Urbana-Champaign

Text Mining Scholarly Publications Using APIs

Abstract

Researchers can analyze their own custom datasets of scholarly publications using text mining. However, digital publications come with copyright licenses, and many are not open access, creating obstacles for researchers such as downloading the full text of the papers. In this report, I describe my work constructing a pipeline that will download full texts of scholarly publications to help researchers create their own custom datasets. My pipeline is reusable such that given any Digital Object Identifier (DOI) of scholarly papers it can extract the papers' PDF and XML full texts, if available, and store them in a database. To extract the full text under various copyright licenses, I used text and data mining APIs supplied by Crossref, Elsevier, and Wiley. I also identified scientific analysis tasks that could be done after the extraction by interviewing researchers who mine and analyze scholarly publications as part of my requirements analysis. The full text extraction pipeline is important because it allows different datasets of scholarly publications to be created using a single pipeline, making it easier for researchers to construct their custom datasets without wasting time on copyright licenses.

Keywords: text mining; scholarly publications; full text; requirements analysis; scholarly data mining; XML

1 Introduction

New research is hardly possible without studying previous research. But with the huge volume of digital publications at their disposal, researchers often create custom datasets for their work instead of using whole corpora of scholarly publications. Some researchers might want to study entire venues; for instance, Bolaños (2022) constructed and analyzed a corpus of 227 papers published in the Bibliometric-enhanced Information Retrieval workshop series.
Other researchers create new datasets from large corpora, such as the InTeReC dataset that contains sets of sentences with in-text references to understand citation contexts in scientific papers (Bertin & Atanassova, 2018). The creation of such datasets is a complex task and requires important skills (Bertin & Atanassova, 2018) not only for acquiring the full text but also for dealing with variations in copyright licenses.

In this report, I describe my work constructing a pipeline that will make the creation of these custom datasets easy. My pipeline will be reusable such that given any Digital Object Identifier (DOI) of scholarly papers it can extract the full texts, if available, and researchers can create their own datasets to analyze the papers. The pipeline extracts the papers' PDF and XML full texts and stores them in a database. I identified scientific analysis tasks that can be done after the extraction such as Citation Context Analysis (Anderson & Lemken, 2023) and Funder Information Extraction (Alexander & Vries, 2021). My data extraction pipeline is important because it allows different datasets to be created using a single pipeline making it easier for researchers to construct their custom datasets.

The remainder of this report is divided into six main sections. Section 2, Background, describes some common terms used in the report. Section 3, Methods, discusses my approach to the data extraction pipeline and the requirements analysis. Section 4 describes the results. Section 5 discusses the limitations and future work. Finally, in section 6, I conclude.

2 Background

2.1 Text Mining

Text mining, defined as the process of finding and analyzing patterns in digital text (Lammey, 2014) has become an important tool for researchers to navigate large datasets (Dai et al., 2010) that often contain information such as full text of scholarly documents. It is the process of finding information and detecting patterns in large datasets (Martinez, 2005). It can be used to process research articles (Raja et. al., 2017) to find meaningful information.

2.2 Scholarly Document Processing

Scholarly document processing uses different resources to gather scholarly documents to process the documents for conducting text analysis via text mining (Salloum et al., 2017). Text analysis can be done via semantic analysis to analyze the document (Salloum et al., 2017). Scholarly document processing can be done manually as well but with the huge disposal of scholarly

publications available, automatic text mining is more efficient for the extraction of information from large documents (Lamurias & Couto, 2019).

2.3 Requirements Analysis

Requirements are part of a project that helps achieve certain aspects of a system. A system is a human-produced object which can be physical or virtual (Meyer, 2022, pp 7). In our case, it is a virtual product. Requirements are a tool for producing systems and not a goal in themselves. Requirements are not prepared with perfection in mind, rather they are created in order to produce quality in the system without spending too much effort on preparing the requirements (Meyer, 2022, pp 7). There are many types of requirements – in this project, our requirement is of the type, *goal*, in which there is a need to overcome obstacles to achieve specific results (Meyer, 2022, pp 38). The *goal* requirement needs enough information to have a basic sketch of the project with minimal use of specialized terms for a wide audience to understand (Meyer, 2022, pp 31).

2.4 Identifiers

A Digital Object Identifier (DOI) is a unique alphanumeric character string that identifies a scholarly publication. It provides a link for the digital publication on the internet.[1] The Open Researcher and Contributor ID (ORCID) is a unique ID for researchers that helps identify their work.[2] It helps in author disambiguation, that is, it helps distinguish between authors with the same names and authors who publish under different names (Tekles and Bornmann, 2020). A PubMed ID (PMID) is a unique identifier for papers indexed in PubMed, assigned by the U.S. National Library of Medicine, which is part of the National Institutes of Health.[3]

2.5 Existing Corpora versus Custom Data Sets

The switch from traditional paper journals to digital scholarly publications (Dai et al., 2010) has enabled the creation of corpora such as PubMed Central[4], PLOS Articles[5], S2ORC[6] (Lo et al., 2020), the Association of Computational Linguistics proceedings[7], and more that contain huge volumes of open access scholarly publications that can be text mined. It's common to process corpora and release an enhanced dataset to facilitate citation context as evident from the OpCitance dataset[8] (Hsiao & Torvik, 2023) and the InTeReC dataset (Bertin & Atanassova, 2018). Researchers who want to analyze specific issues want custom datasets instead of looking at the huge existing corpora so that they can get specific results. This is because custom datasets allow researchers to focus on a specific subject that they want to analyze without getting lost in the huge existing corpora that contain millions of scholarly papers.

---

[1] https://www.doi.org
[2] https://info.orcid.org/what-is-orcid/
[3] https://nexus.od.nih.gov/all/2015/08/31/pmid-vs-pmcid-whats-the-difference/
[4] https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/
[5] https://plos.org/text-and-data-mining/
[6] https://github.com/allenai/s2orcCite
[7] https://github.com/acl-org/acl-anthology/
[8] https://doi.org/10.13012/B2IDB-4353270_V2

2.6 Copyright vs. Open Access

Copyright licenses are authorizations given by the author to the publisher to sell, distribute and publish their work.[9] Digital publications come with copyright licenses, and many are not open access, creating obstacles for researchers (Molloy et. al., 2016).

2.7 Existing Data Sources

To collect license information, I use Crossref, an association that has its own Text and Data Mining (TDM) Application Programming Interface (API) (Polischuk, 2020), that provides metadata for scholarly publications from their DOIs. However, if the publications are licensed by Elsevier or Wiley, Crossref cannot provide the full texts (Vikery, 2021), therefore, I need to use those publishers' TDM APIs to download the full texts.

3 Methods

3.1 Data Extraction Pipeline

My data extraction pipeline extracts the full text of scholarly publications using the Crossref[10], Elsevier[11], and Wiley[12] APIs.

First, the pipeline gets the DOI of the scholarly paper from the user via a text file[13] containing one DOI on each line and supplies it to the Crossref TDM API which returns the metadata of the paper. The pipeline then searches for the license URL and full-text URL. Then, if available, it downloads the PDF and XML versions of the full text and stores the information in a database as shown in Figure 1.

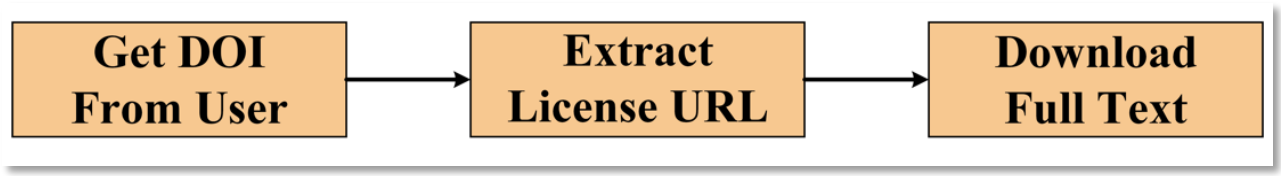

Figure 1: Data Extraction Pipeline

Since some publishers have copyright licenses that are not accessible via Crossref (Vickery 2021), users cannot access full texts from these publishers but must instead use the publisher's own APIs to access the full text. Therefore, my pipeline checks whether the publisher is Wiley or Elsevier and if so, accesses the full text from their respective TDM APIs.

The pipeline also supplies an output file[14] that gives an analysis for each DOI about what license URL there is and if the full text was downloaded or not.

---

[9] https://www.legalzoom.com/articles/what-is-a-copyright-license
[10] https://www.crossref.org/documentation/retrieve-metadata/rest-api/text-and-data-mining/
[11] https://www.elsevier.com/about/policies/text-and-data-mining
[12] https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining
[13] https://github.com/infoqualitylab/text-mining-scholarly-API/blob/main/heng_dois.txt
[14] https://github.com/infoqualitylab/text-mining-scholarly-API/blob/main/output.txt

3.2 Requirements Analysis

In my project, I conducted requirements analysis to identify applications of full text datasets, which I also call scientific analysis tasks. I interviewed researchers who mine and analyze scholarly publications to understand what they need in analyzing automated custom datasets. I spoke to 4 people who were identified by my mentor as working on custom datasets to analyze specific problems. I identified use cases for the full text datasets to implement their work in my pipeline.

4 Results

4.1 Results of the Requirements Analysis

After the 4 interviews, there were 4 possible projects to work on. The first interviewee (1) was a post-doctoral researcher working on Citation Context Analysis and the second (2) was a Ph.D. student working on Funder Information Extraction, as shown in Figure 2. The third interviewee (3) was a Ph.D. student who was also interested in full text and Citation Context Analysis; her work deals with biomedical literature in the PubMed database, however, I did not have enough time to work on this project. The fourth interviewee (4) was a Ph.D. student who analyzed very old scholarly papers that were not digitally published, they were scanned and uploaded to databases which made their XML versions unavailable, and they often did not have DOIs either, which made it difficult for me to work on this project. However, it was interesting to find out that people have such various needs with these custom datasets.

From Figure 2, we can see that I identified two possible applications for XML full text: Funder Information Extraction and Citation Context Analysis, which could support ongoing projects conducted by other researchers.

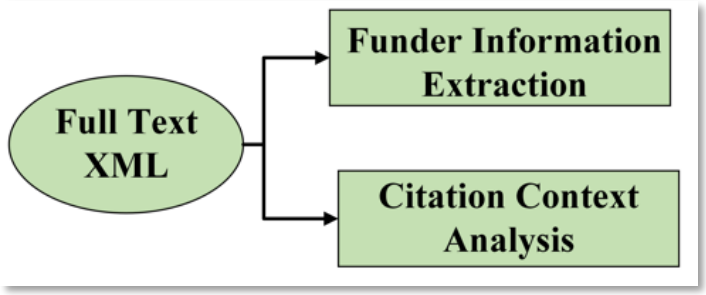

Figure 2: Possible Applications of Full Text XML

4.1.1 Funder Information Extraction

In this project, I parse the XML full text to extract the acknowledgment section as shown in Figure 3, and along with that, the authors' information such as full name, ORCID ID, and affiliation, DOI, year published, and type of paper. This creates a dataset with funding information for many papers using their DOIs. This information is crucial to analyze the funder's relationship with the research being conducted to detect funding bias.

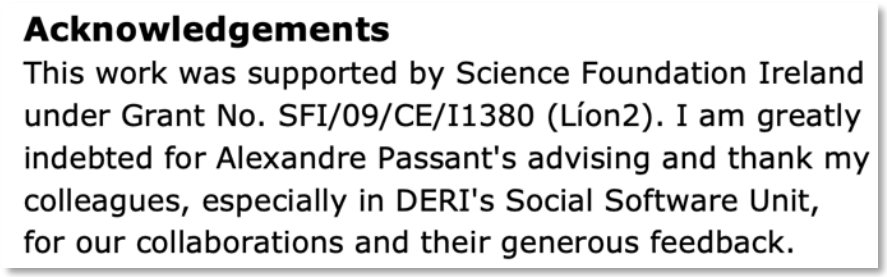

Figure 3: Acknowledgment Section of a Scholarly Publication

4.1.2. Citation Context Analysis

I seek to support two related citation context analysis projects. Project 1: Computational Chemistry Protocol Glitch Project is to extract sentences that occur before and after an in-text citation as shown in Figure 4 along with the reference information to analyze the citation context. By extracting the sentences, I create a dataset that helps check the citation contexts of many scholarly papers with their references to check the impact of using a retracted paper on the scholarly papers.

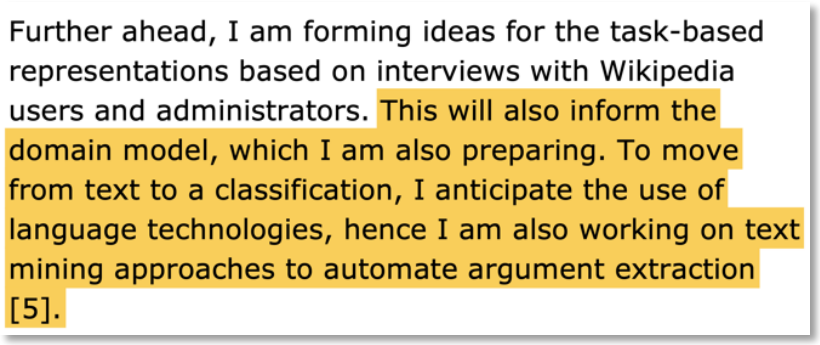

Figure 4: Sentences Around In-Text Citation

The goal of the Biomedical Document Processing Project is to extend the data extraction pipeline to download full text using PubMed IDs instead of DOIs and then use the full text to check for citation errors. This application will also involve extracting sentences around in-text citations as shown in Figure 4 and then checking their context in relation to the reference to verify whether the meaning of the citation has been retained in the paper citing it.

4.2 Results from the Initial Implementation of Citation Context Analysis

4.2.1 Computational Chemistry Protocol Glitch Project

From Figure 5, we can see the output for a dataset with 287 DOIs[15] that all cite a paper that had (Willoughby et al., 2014) a code glitch. Out of the 287 DOIs, the pipeline could only download

[15] https://github.com/infoqualitylab/text-mining-scholarly-API/blob/main/heng_dois.txt

87 DOIs[16] (30%) that had valid full texts in plain text, PDF, and XML formats. The rest 200 DOIs (70%) had errors of various types. 62 DOIs (31%) did not have license information in the metadata, 81 DOIs (40.5%) only had "unspecified" type of URLs which meant they could not be saved in PDF, XML, or plain text formats. 57 DOIs (28.5%) got downloaded but did not have full text for 3 reasons. The first is that an invalid XML file was downloaded for 1 DOI because the university did not have access to that article. 19 PDF and XML files each were downloaded for 19 DOIs, out of which 18 DOIs were under a Wiley license which could not be accessed, and 1 DOI was published by Sage Journals and could not be accessed. 37 PDF files were downloaded for 37 DOIs, out of which 1 DOI could not be accessed because the university did not have access to it, 12 DOIs were under a Wiley license which could not be accessed, and 24 DOIs were published by American Chemical Society (ACS) and could not be accessed because ACS sells its TDM services which means to download full text from ACS one would have to purchase their TDM service to gain access to the API token that would give permission to allow full text downloads[17].

[16] https://github.com/infoqualitylab/text-mining-scholarly-API/blob/main/output.txt
[17] https://solutions.acs.org/solutions/text-and-data-mining/

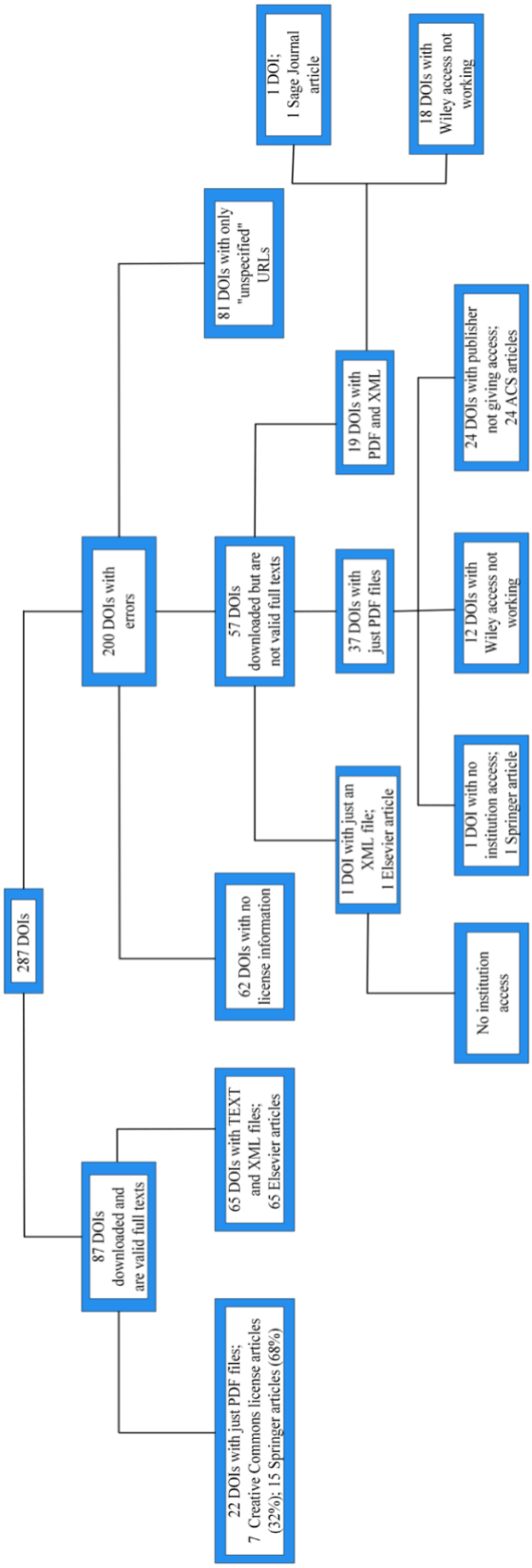

Figure 5: Output After Passing the Pipeline for 287 DOIs

5 Discussion

My pipeline aims to remove difficulties for researchers to create and analyze custom datasets. The analysis pipelines help the analysis to be done automatically. It is important to conduct more test cases to understand how many full text papers can actually be extracted by testing DOIs from other researchers' custom datasets. It will be helpful to know whether other researchers who text mine scholarly publications had similar problems with creating and analyzing custom datasets and how they dealt with them.

The implication of the results from Computational Chemistry Protocol Glitch Project (Section 4.2.1) was that even though the institution has licenses for many scholarly publications, it was not always possible to text mine the PDF and XML full texts of those papers using the current data extraction pipeline. Even though the PDF could be downloaded manually, it gave access issues when being accessed through an API. The XML files, however, are not easily available to download manually even when the document is under the institution's license, thus, usually the only way to get access to the XML file is via the TDM API. The data extraction pipeline brought these issues to light and will continue to bring other important problems related to the text mining of scholarly documents to light by working on other projects as identified in the requirements analysis and continuing to identify more projects.

5.1 Limitations

The pipeline currently does not download full text from all given DOIs as seen from the above results in Computational Chemistry Protocol Glitch Project (Section 4.2.1). The pipeline needs to identify whether a PDF or XML file is valid and contains full text that can be text mined. Additionally, it needs to check whether or not a given DOI can be accessed by that institution so that it can inform the user whether or not the full text can be downloaded. Finally, the pipeline should also be able to tell the user whether the publisher has a separate paid subscription for its TDM service to access the full text.

5.2 Future Work

I will continue to work on developing my pipeline to implement these applications. Currently, I have identified 3 future projects. The first project, *Valid full text XML Checker*, checks whether or not a downloaded XML file is a text minable XML file with full text. The second project, *Valid full text PDF Checker*, checks whether or not a downloaded PDF file is a text minable PDF file with full text. The third project, *Institution Access Checker*, checks whether or not a given DOI is accessible within that institution's subscriptions for downloading full text. With further requirement analysis, I will investigate and incorporate additional possible applications for analyzing full texts of various scholarly publications.

6 Conclusion

I have constructed a pipeline that will make it easy for researchers to create their custom collection of data for research using DOIs and Crossref, Elsevier, and Wiley's TDM APIs. My pipeline will help navigate the license problems and other access issues related to full text

extraction and allow researchers to focus on their analysis work. Various applications will be able to analyze the full texts already acquired from the pipeline.

7 Code Availability

My code for the pipeline can be found on this GitHub URL: https://github.com/infoqualitylab/text-mining-scholarly-API.

8 Outputs

1. Sarraf, I. (June 16, 2023). *DREU Project Planning Presentation.* Presented at the Information Quality Lab. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1235242303738?s=6u28zaxdhxws20bazermdbs0bzfev7vy
2. Sarraf, I. (July 6, 2023). *Text mining scholarly publications using APIs.* Accepted to the METSTI 2023 Workshop (ASIS&T, 2023). https://uofi.app.box.com/file/1254524512648?s=oett84cyop8n11wtqybq7f0lb0z48ub4
3. Sarraf, I. (July 25, 2023). *Text mining scholarly publications using APIs*. Presented at the Information Quality Lab. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. https://hdl.handle.net/2142/120049
4. Sarraf, I. (July 26, 2023). *Text mining scholarly publications using APIs*. Presented at the Illinois Summer Research Program Alliance STEM Career Exploration and Symposium. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. https://hdl.handle.net/2142/120049
5. Sarraf, I. (August 1, 2023). *Heng's dataset analysis.* Presented at the Information Quality Lab. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1271502491606
6. Sarraf, I. (August 2, 2023). *Text mining scholarly publications using APIs.* Presented at the iSchool Colloquium. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1266500361465?s=r8z03z5na1w3m82licy3ze8kewd4ektx
7. Sarraf, I. (August 4, 2023). *Code Availability.* University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available at https://github.com/infoqualitylab/text-mining-scholarly-API
8. Sarraf, I. (August 4, 2023). *Valid full text XML Checker* [Future project]. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1273092175164
9. Sarraf, I. (August 4, 2023). *Valid full text PDF Checker* [Future project]. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1273089810354
10. Sarraf, I. (August 4, 2023). *Institution Access Checker* [Future project]. University of Illinois at Urbana Champaign, Urbana-Champaign, Illinois. Available privately to members of the Information Quality Lab at https://uofi.app.box.com/file/1273093885774

9 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). I would like to thank my mentors Dr. Jodi Schneider and Ph.D. mentor Yuanxi Fu who were supported by NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis for their constant guidance. Additionally, I would like to thank my family and friends for supporting me throughout this research process.

10 References

Alexander, D., & Vries, A. P. de. (2021). “This research is funded by...”: named entity recognition of financial information in research papers. In I. Frommholz, P. Mayr, G. Cabanac, & S. Verberne (Eds.), *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval* (Vol. 2847, pp. 102–110). CEUR. https://ceur-ws.org/Vol-2847/paper-10.pdf

Anderson, M. H., & Lemken R. K. (2023). Citation context analysis as a method for conducting rigorous and impactful literature reviews. *Organizational Research Methods, 26*(1), 77-106. https://doi.org/10.1177/1094428120969905

ASIS&T. (2023, July 19). *METSTI 2023 Workshop CFP*. Retrieved August 1, 2023, from https://www.asist.org/sig/sigmet/news/#

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

Bolaños, F. (2022). Mapping the trending topics of bibliometric-enhanced information retrieval. In I. Frommholz, P. Mayr, G. Cabanac, & S. Verberne (Eds.), *Proceedings of the 12th International Workshop on Bibliometric-enhanced Information Retrieval* (Vol. 3230, pp. 61–70). CEUR. https://ceur-ws.org/Vol-3230/paper-08.pdf

Dai, HJ., Chang, YC., Tsai, R. TH., Hsu, WL. (2010). New Challenges for Biological Text-Mining in the Next Decade. *Journal of Computer Science and Technology, 25(1)*, 169–179. https://doi.org/10.1007/s11390-010-9313-5

Hsiao, T.-K., & Torvik, V. I. (2023). OpCitance: Citation contexts identified from the PubMed Central open access articles. *Scientific Data, 10*, 243. https://doi.org/10.1038/s41597-023-02134-x

Lammey, R. (2014). Crossref’s text and data mining services. *Learned Publishing, 27(4),* 245-250. https://doi.org/10.1087/20140402

Lamurias, A., Couto F. M. (2019). Text Mining for Bioinformatics using biomedical literature. *Encyclopedia of Bioinformatics and Computational Biology, 1,* 602-611. https://doi.org/10.1016/B978-0-12-809633-8.20409-3

Lo, K., Wang, L. L., Neumann, M., Kinney, R., Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.447

Meyer, B. (2022). Requirements: basic concepts and definitions. In *Handbook of Requirements and Business Analysis* (pp. 1-20). Springer, Cham. https://doi.org/10.1007/978-3-031-06739-6_1

Meyer, B. (2022). Requirements: general principles. In *Handbook of Requirements and Business Analysis* (pp. 21-34). Springer, Cham. https://doi.org/10.1007/978-3-031-06739-6_2

Meyer, B. (2022). Standard Plan for requirements. In *Handbook of Requirements and Business Analysis* (pp. 35-46). Springer, Cham. https://doi.org/10.1007/978-3-031-06739-6_3

Molloy, J., Haeussler, M., Murray-Rust, P., Oppenheim, C. (2016). Working with text: tools, techniques, and approaches for text mining. *Chandos Information Professional Series*. https://doi.org/10.1016/B978-1-84334-749-1.00004-4

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/

Raja, K., Patrick, M., Gao, Y., Madu, D., Yang, Y., Tsoi, L. C. (2017). A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. *International Journal of Genomics*, pages 1-10. https://doi.org/10.1155/2017/6213474

Tekles, A., Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies 2020.1(4),* 1510–1528. https://doi.org/10.1162/qss_a_00081

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/

11 Appendix

(1) Heng Zheng
(2) Malik Oyewale Salami
(3) Maria Janina De La Cruz Sarol
(4) Yuerong Hu