The Computational Literature Review package (CLR2): a user guide

Michael Mortenson
Warwick Business School, UK

Richard Vidgen
UNSW Business School, Australia

1. Introduction - conducting a research project

The starting point of your project is to identify a research area that interests you (item 1 in Figure 1). For example, this might be "the impact of AI in education". You should consider why this is important (the "so what" question) and what you broadly want to achieve through the research. Of course, you may not know this when starting out but as you navigate through the literature then your motivation and the importance of your proposed research should become clearer.

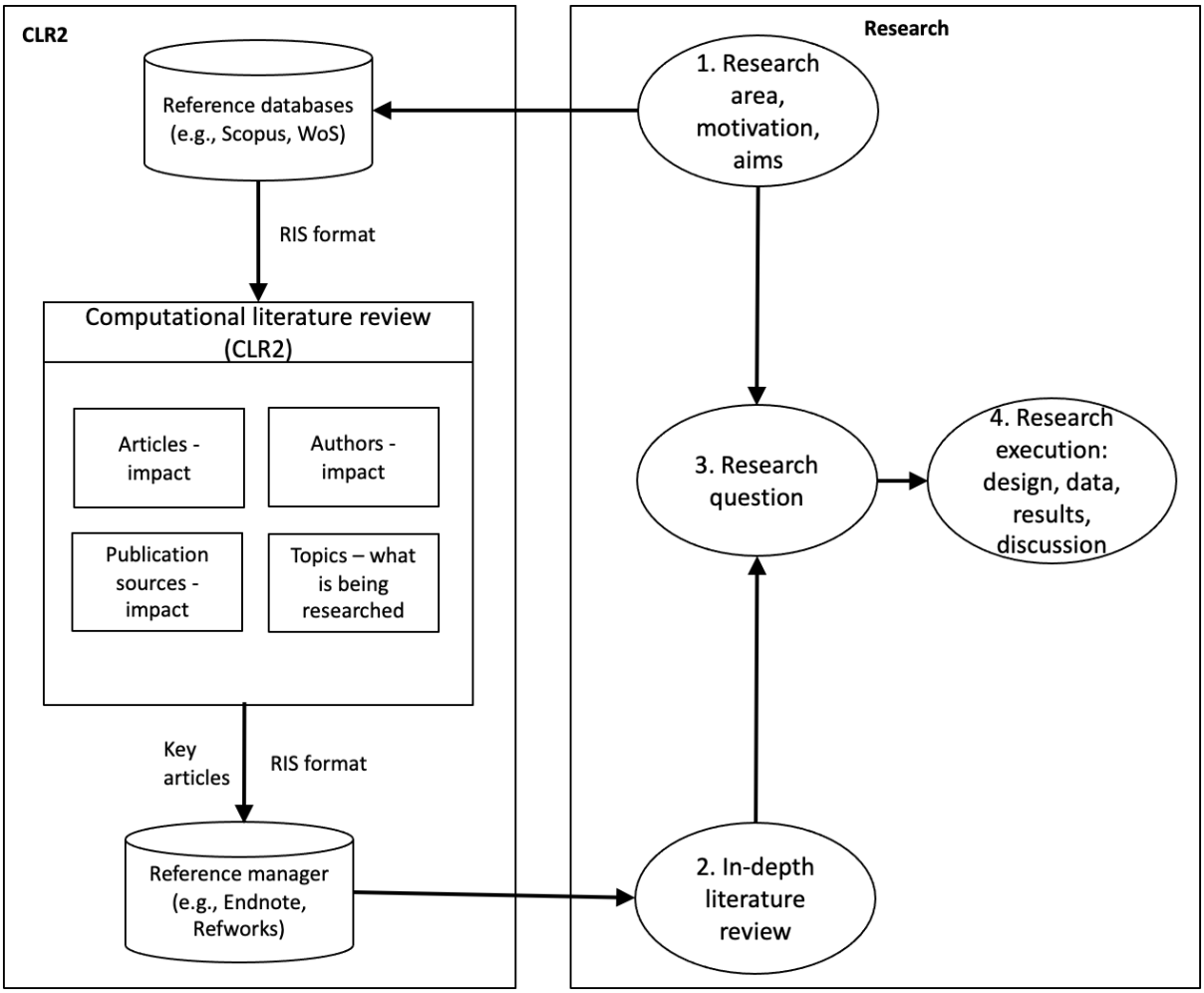


Figure 1: The research process with CLR2 support

To find out what is going on in AI and education we will typically access a citation database such as Scopus or Web of Science. Your initial query might return thousands of records – far too many for you to make sense of, let alone attempt to review manually. To overcome the limitations of traditional and systematic literature reviews, researchers have turned to computational literature reviews (CLRs), which leverage machine learning algorithms to analyse large volumes of literature (Vidgen and Mortenson, 2016). Antons et al. (2023) define a CLR as: "a structured process intended to augment human researchers' information processing capabilities through the use of machine learning algorithms that help analyse the content of a comprehensive text corpus in a specific knowledge domain (e.g., a research topic, academic journal or scientific field) in a way that is scalable and real-time capable." (p.109).

The CLR2 software (the lefthand side of Figure 1) will help you make sense of large volumes of literature and give you insight into the research topics that are driving the content of your search results (for example, what are researchers working on that has given rise to the body of knowledge in the AI and education field?). As well as using the CLR2 to make sense of a corpus of literature, you also need to do an in-depth review to see what has been done before, where the research gaps might be, and to refine your research area into a specific and answerable research question (Figure 1, item 3).

According to Tranfield et al. (2003), “the aim of conducting a literature review is often to enable the researcher both to map and to assess the existing intellectual territory, and to specify a research question to develop the existing body of knowledge further.” (p. 208). Literature reviews also contextualise researchers’ work within the broader scholarly discourse and enable the adoption of appropriate theoretical frameworks, research methods, and data analysis techniques (Kitchenham and Charters, 2007). However, traditional literature reviews have limitations. They rely on subjective selection processes, which may introduce bias and limit the comprehensiveness of the review (Hart, 1988) and run the risk of being idiosyncratic, incomplete, and overly subjective, with the review author(s) being “free to pick and choose” the papers to include (and exclude) (White and Schmidt, 2005, p. 54).

CLRs address the weaknesses of traditional reviews by reducing reliance on subjective selection processes and minimising bias introduced by human judgement (Antons et al., 2023; Mortenson and Vidgen, 2016). CLRs augment researchers’ information processing capabilities by leveraging machine learning algorithms to analyse the content of papers, thereby providing a more comprehensive assessment of the literature (Antons et al., 2023; Leeflang and Wittink, 2000).

However, the CLR approach does not obviate the need to read, review, and synthesise the literature relevant to your research project. Your detailed review (item 2 in Figure 1) should encompass a number of articles that is feasible for human review (we suggest a range of 30 to 50 articles). These will be included in the reference section of your project report or dissertation.

Inspection of the CLR2 output will help you to refine your research area and to focus in on a specific research question. The process of aligning the research area with a research question and an in-depth literature review is iterative – i.e., it may take a few cycles to settle. Using the CLR2 software accelerates this process, allowing you to get a broad view of a research area while supporting drilling down into more focussed areas to arrive at a research question that is clear, specific, and answerable. Once the research question is stable then the research design and data collection can take place.

For example, a review of the topics identified from the AI and education search might lead you to focus in on questions such as “how does AI impact on pedagogy?”, or “What are the barriers and enablers to implementing AI in education”, or “What are the benefits and risk of AI in education”. I.e., we need to move from a general area of interest (item 1 in Figure 1) to a specific research question (item 3 in Figure 1). We recommend you have a single overarching research question, which might be broken down into sub-questions that together will enable you to answer the larger question. Having a single question provides focus to your research; if you address it effectively then you are more likely to make a useful contribution to theory and to practice. Arriving at your research question will likely involve an interative process in which the search string is modified and the CLR2 rerun, results inspected, until the question settles and the relevant literature has been identified.

Once the research question is settled and the literature review complete then the research design can take place (item 4 in Figure 1), i.e., how will you answer the research question? This might involve an experiment, a survey, interviews, and indeed it might be a literature review (if the method is to conduct a literature review then likely a larger number of articles will be needed and there will be greater analytical depth of the content of the articles).

2. CLR2 installation

The CLR2 is a redevelopment of the original CLR proposed and built by Mortenson and Vidgen (2016). The aim of the CLR2 is (1) to show articles, authors and publication sources have the most impact, (2) provide insight into the topics of research that underlie the corpus of research being examined, and (3) to produce an output file that can be used as the basis for an in-depth literature review. The CLR2 accepts input in RIS format and produces output in RIS format, thus enabling data to be exported and imported from any reference manager software (e.g., EndNote).

The CLR2 is developed in Python and is most easily accessed using the Google Colab environment: https://colab.research.google.com/. The benefits of Colab are that it allows you to write and execute Python in your browser, with zero configuration required, access to GPUs free of charge, and easy sharing of code. Perhaps most importantly, no programming knowledge is needed to execute a Colab workbook, such as CLR2.

To access CLR2 on Colab you will need a Google account and an associated Google drive to store your copy of the workbook.

Access the CLR2 using this link:

https://colab.research.google.com/github/MJMortensonWarwick/computational_lit_review/blob/main/CLR_demo.ipynb

Once you have loaded the workbook (Figure 2), make a copy by selecting “File” and then “Make a copy in Drive”. This will give you a local copy that you can edit and save.

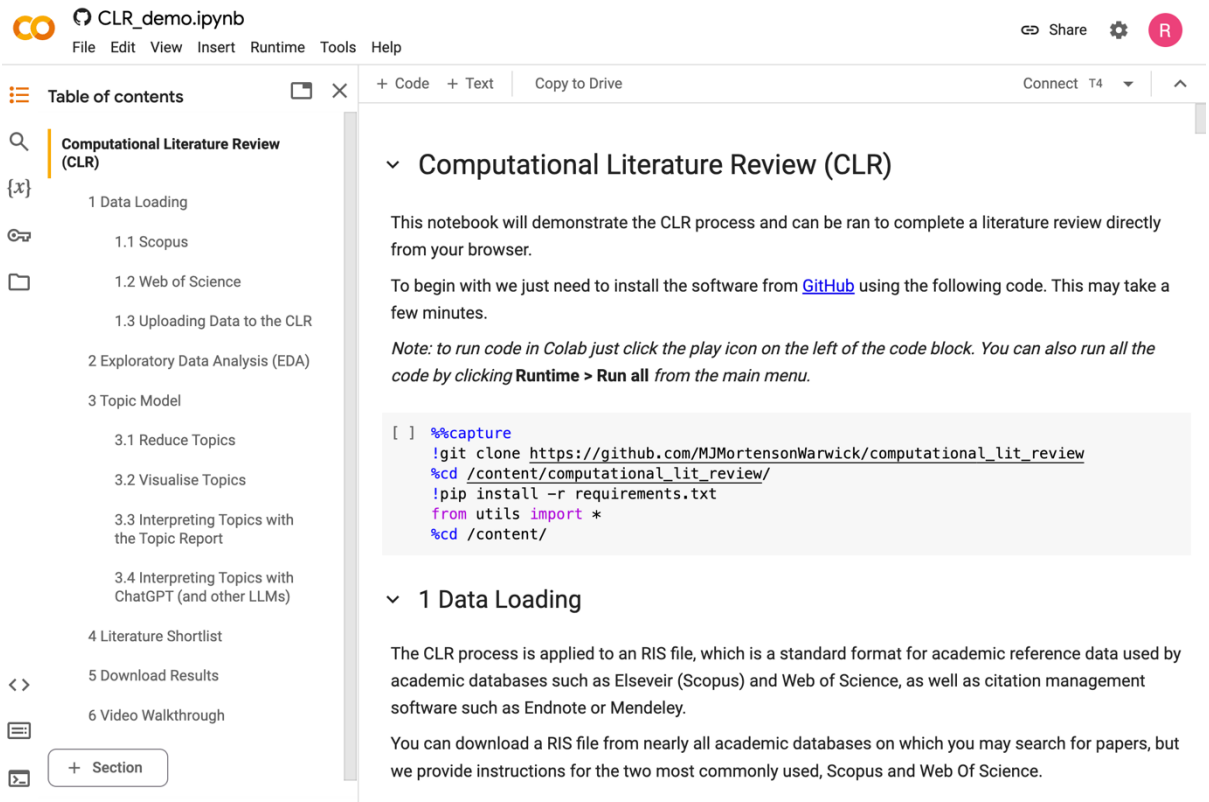


Figure 2: CLR2 in Colab

3. Data collection and preparation

Data for the corpus of articles is retrieved from Scopus, which claims to "deliver more global content (50%-230% more depending on region) than the nearest competitor" (https://www.elsevier.com/solutions/scopus/why-choose-scopus). Although Scopus is widely recognised as a high-quality data source for literature reviews, there are alternatives, such as Thomson Reuters Web of Science (WoS). Singh et al. (2021) compared Scopus and WoS and found that WoS is more selective than Scopus and that 99.11% of the journals indexed in WoS are also indexed in Scopus.

You will need to export a file in RIS format from your referencing software, e.g., Scopus, or Web of Science. Any referencing software that supports RIS export can be used to create your source data for CLR2.

As Scopus is widely used and has broad coverage of the literature, we provide an illustration here. For example, assume that you are interested in AI and education and create a Scopus search (Figure 3):

(TITLE-ABS-KEY ("artificial intelligence") AND TITLE (education*))

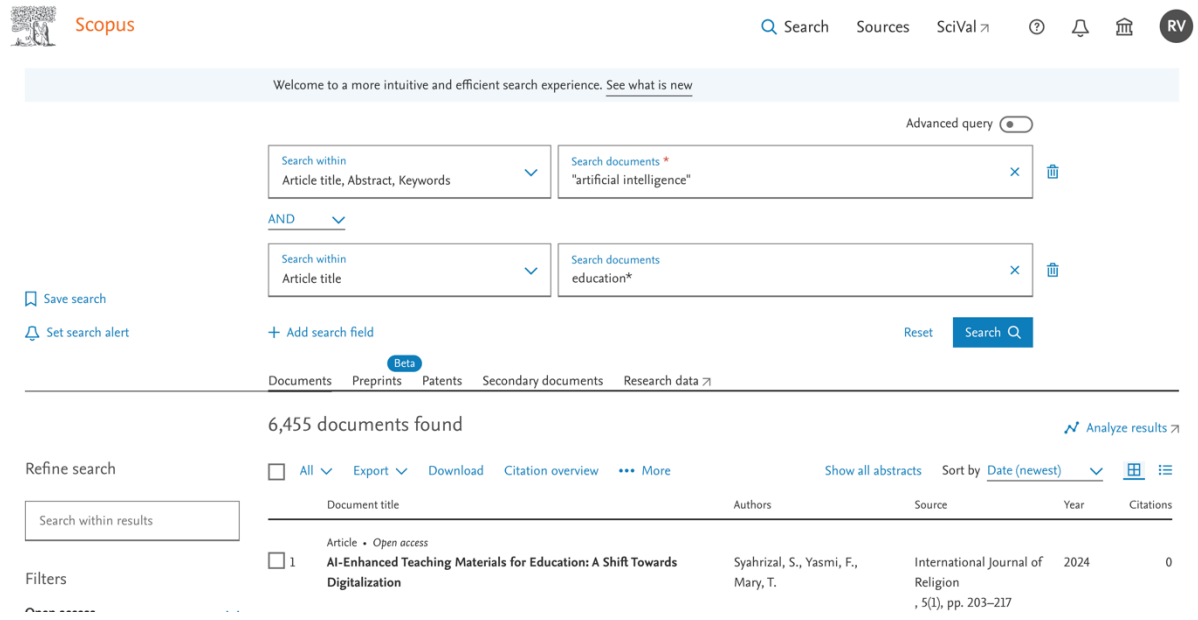


Figure 3: search in Scopus

This search returns 6,455 documents (as of 31 March 2024). Now we want to export these document references and abstracts as a RIS format file. To do this we click the "All" option and then in the "Export" dropdown we select File type RIS. Makes sure that "Citation information", "Bibliographic information", and "Abstract & keywords" are selected (see Figure 4).

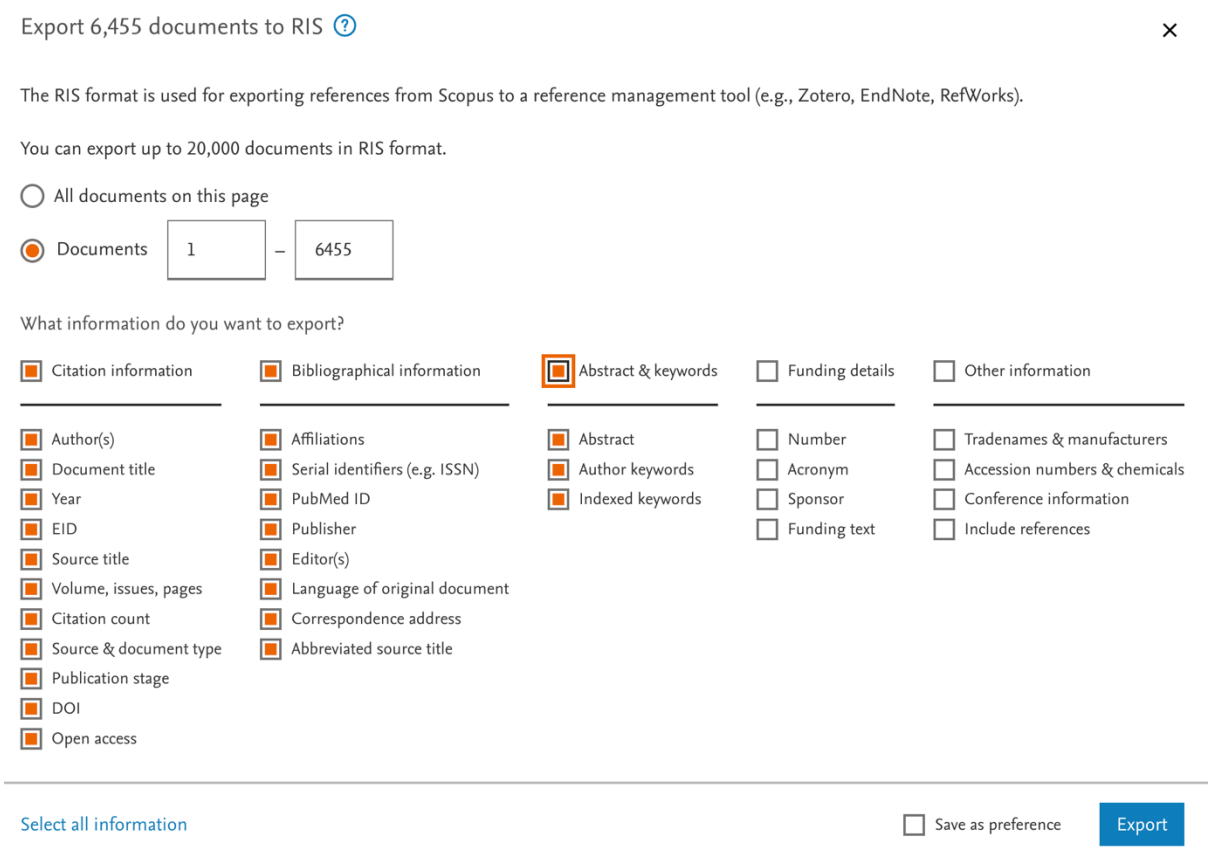


Figure 4: export to RIS in Scopus

Once you click on “Export” the download will start and you will then have a RIS format file that can be uploaded to the CLR2 Colab notebook. If you need more than 20,000 records then the search will need to be divided into sub-searches, for example based on publication year.

If your institution uses Web of Science rather than Scopus then the process is broadly the same, i.e., create a search string, select all the articles in the returned results and export them in RIS format for input to the CLR2 (Figure 5).

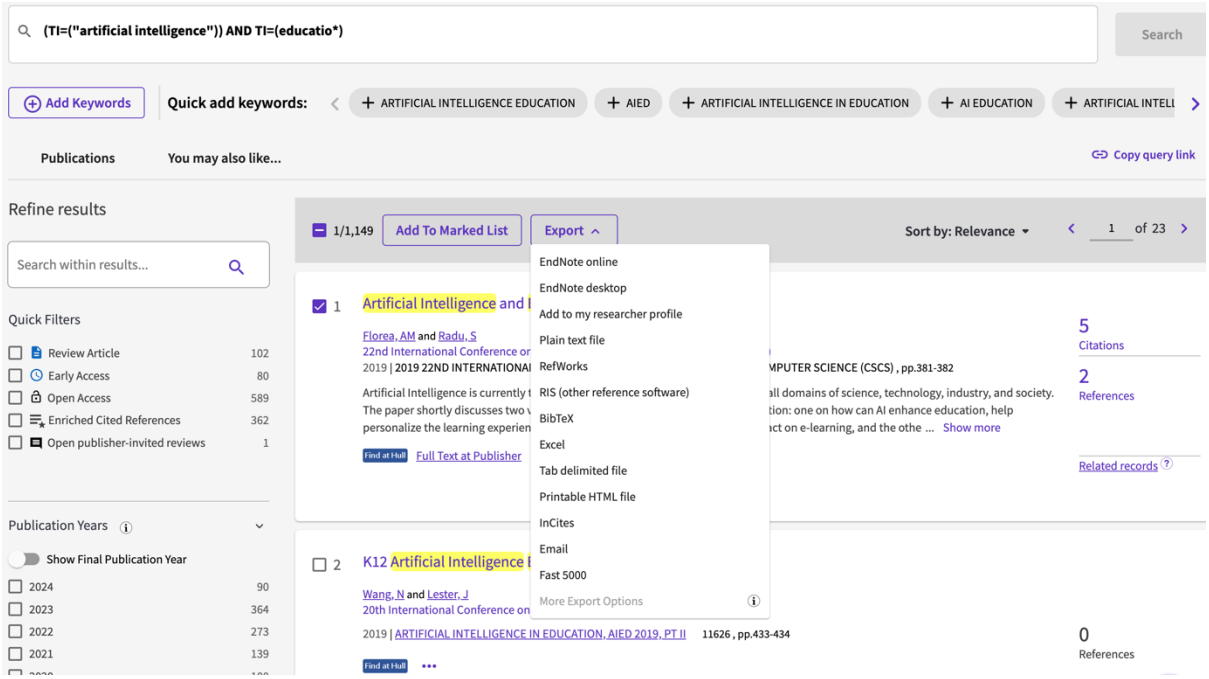


Figure 5: export to RIS in Web of Science

4. Running the CLR2

The simplest way to start with the CLR2 is to run it in standard form using the default options. To do this select “Runtime” and “Run all” in Colab (Figure 2). The only user action required is to tell the CLR2 notebook where the input RIS file sits (this could be in your Google drive, the cloud, or locally on your laptop). This file should be named “scopus.ris” (if you want to have a different file name then you can modify the Colab workbook accordingly).

If you want to see each code section running then you can click on the run button for each section. This approach is required if you want to customise the CLR2, for example to run a slower but more comprehensive model (see Section 6).

5. Interpreting the CLR2 output

The CLR2 produces three types of output:

- Bibliographic information (e.g., which papers have the most citations, which publication sources are most influential?).
- Topic analysis
- Output RIS file for use as part of a detailed literature review

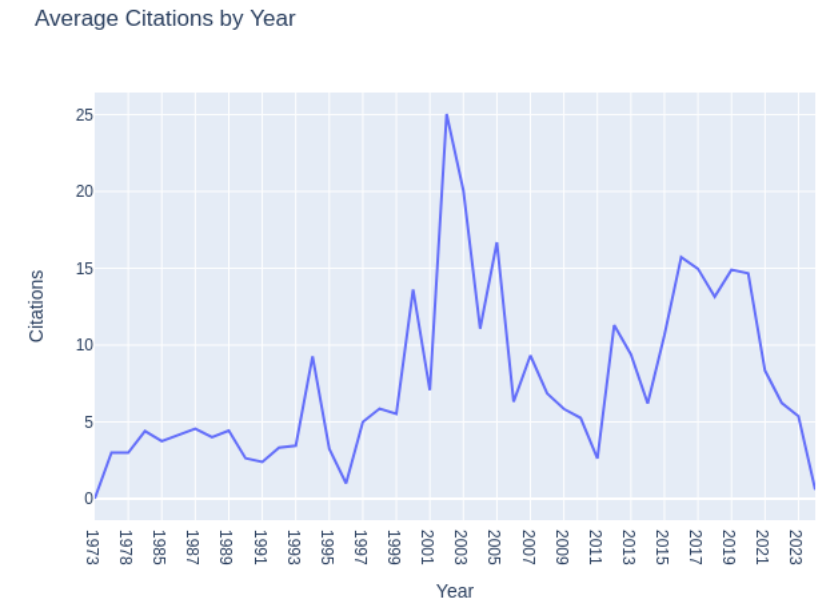
Bibliographic information

Here we are concerned with the impact of articles (which have the most citations?), the impact of authors (which have the most citations and highest h-index in the downloaded

dataset) and the impact of publication sources (which publications have the most citations and highest h-index). This background helps us to begin to recognise key papers, key researchers, and key publications in our chosen research area and thus supports a complete literature review that is less likely to have significant omissions.

The results of the bibliographic are downloaded by the CLR2 in a single file – eda.zip.

The citations per year and number of publications per year give a sense of how your research area is trending. In Figure 6 we can see clearly that the AI and education area took off in 2017 with around 1650 articles in 2023 (the drop in 2024 is due to the timing of the search, i.e., conducted in April).



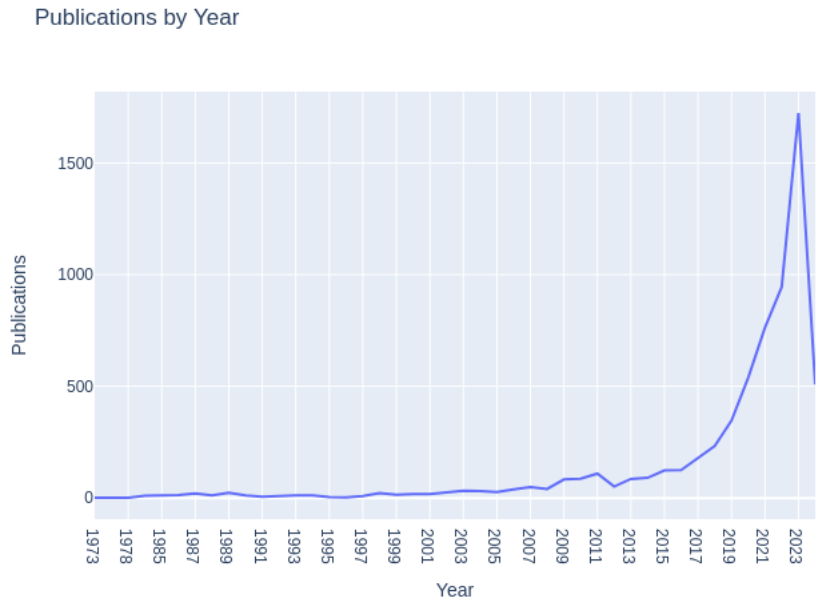


Figure 6: Citations and publications per year

Figure 7 shows the most active authors, by absolute number of citations and by h-index. Note that authors are not unique in systems such as Scopus, e.g., there might be multiple authors named Jesse Smith and a single author might be also listed as Jesse A. Smith. Accordingly, author analysis needs to be approached with caution. It is also worth noting that the h-index we calculate in the CLR2 is local to the search dataset, i.e., it will not likely be the same as the author's h-index on Scopus, and will change as the search string is modified. What is important here is identifying the authors that are most relevant to your research area.

Raw citation counts are useful but it favours authors with one or two big hit papers that have achieved high citations. The h-index balances productivity (number of papers) and citation inpact. It is the largest number, h, such that h articles have at least h citations each. For example, if an author has five publications, with 9, 7, 6, 2, and 1 citations (ordered from greatest to least), then the author's h-index is 3, because the author has three publications with 3 or more citations. However, the author does not have four publications with 4 or more citations (https://en.wikipedia.org/wiki/H-index). Thus, an author with only one paper that has 1,000 citations has ab h-index of 1.

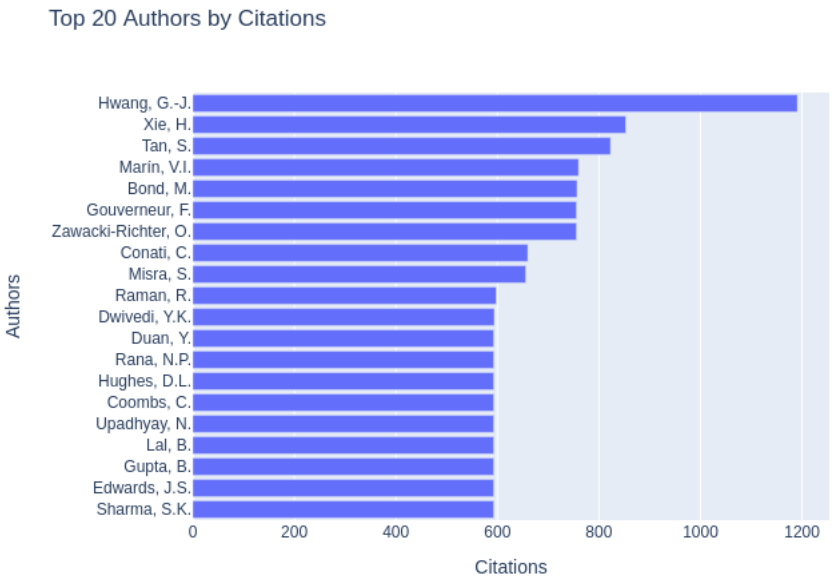


Figure 7: Citations and h-index by author

In Figure 8 we simply the sort the dataset by citation count to see which articles are most impactful. The most highly cited article is a systematic literature review of research on AI in education.

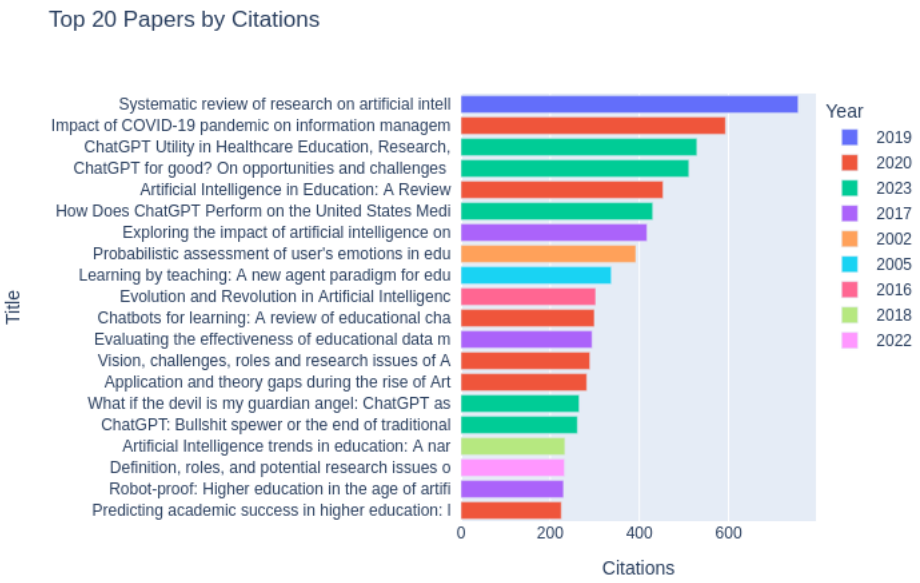


Figure 8: Top articles by citation count

In Figure 9 we show the impact of the different publication sources (e.g., journals, conference proceedings). As with authors, we show the total citations and the h-index (again, calculated locally to our dataset). This gives an idea of which sources are most impactful and helps get a sense of where that research is coming from (e.g., computer sicence, social science,

education). Perhaps unsurprisingly, the most impactful source (for citations and h-index) is "Computers and Education: Artificial Intelligence".

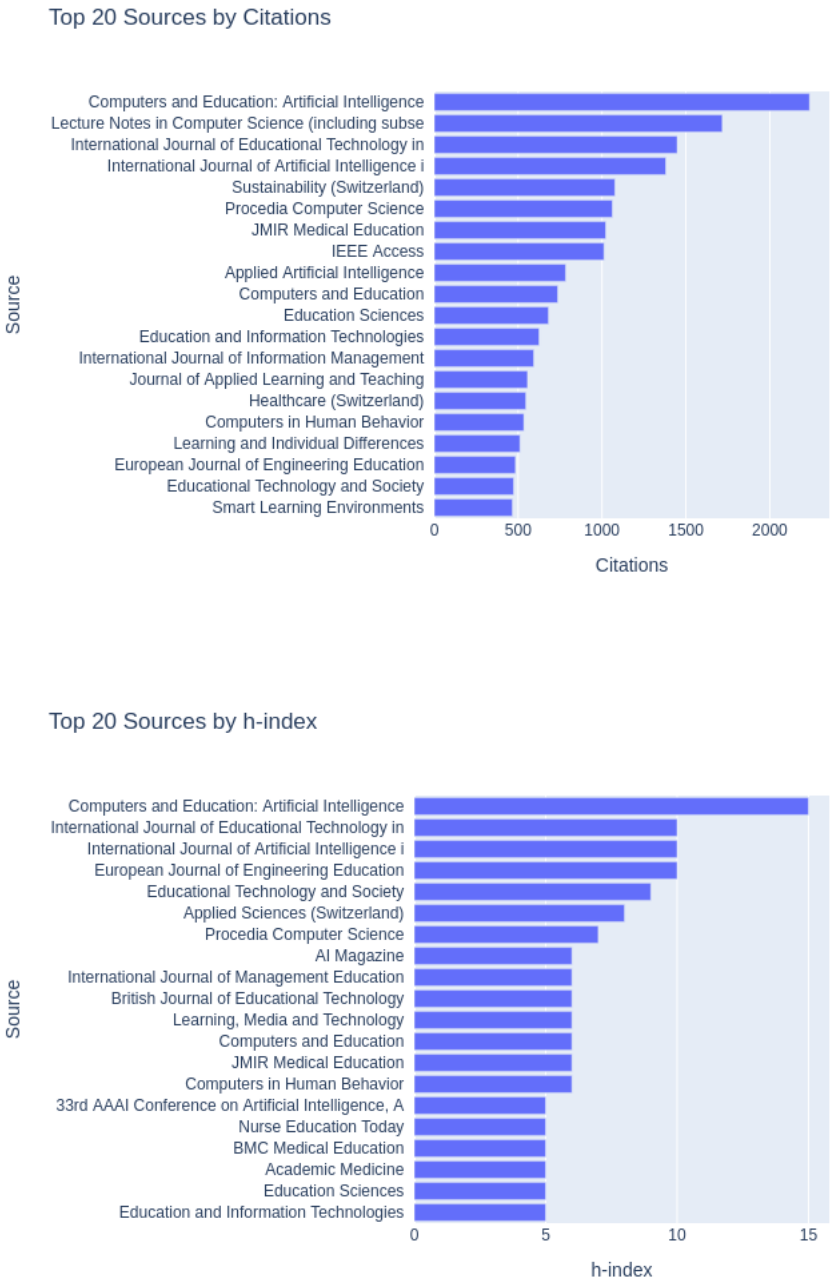


Figure 9: Citations an h-index by publication source

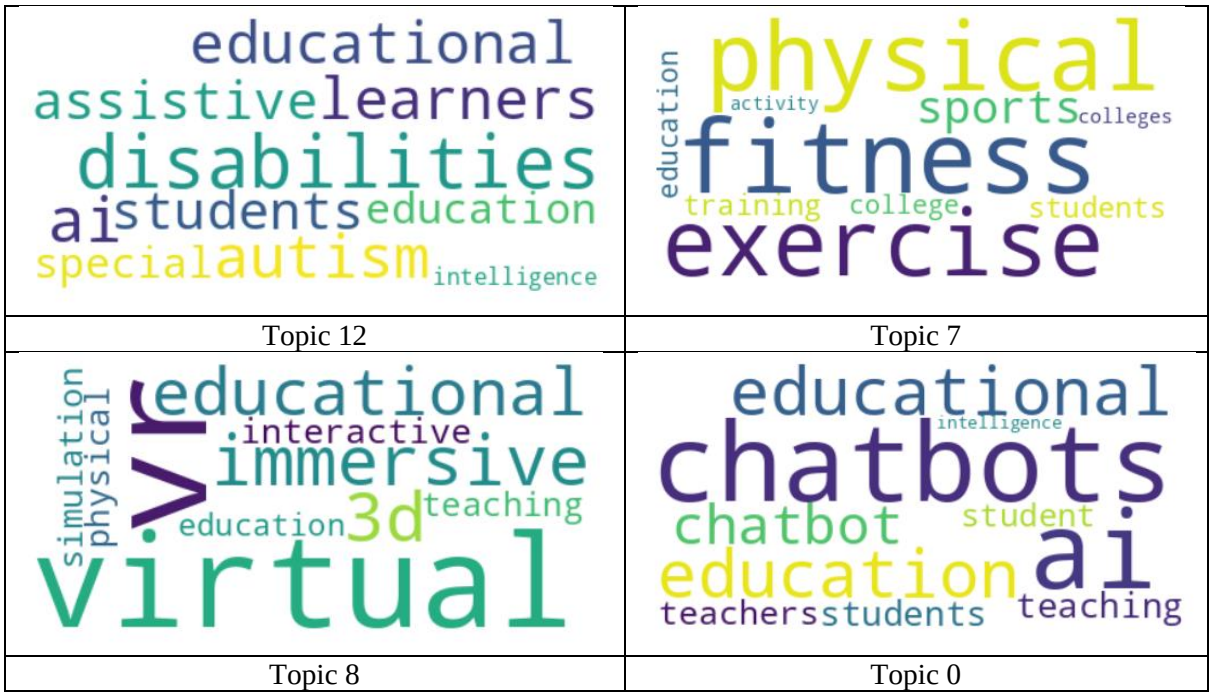
Topic identification

The CLR2 software uses the Bertopic package to implement topic modelling. The results of the bibliographic are downloaded by the CLR2 in a single file – output.zip.

The number of topics is decided automatically and topics that are not well defined are dropped automatically (this can be changed as described in Section 6). The CLR2 has identified 19 topics (0 through 18) together with the top five keywords for each topic (Figure 10). While the CLR2 does not apply labels automatically to the topics, inspection of the keywords provides a good idea of the topic, e.g., Topic 0 appears to be concerned with chatbots and topic 1 with medical applications.

Figure 10: Topics

Each topic has a word cloud to help make sense of, and label, the topic content. In Figure 11 we show sample output for four topics.



Topic 12	Topic 7
Topic 8	Topic 0

Figure 11: Sample word clouds

Topics are interrelated and in Figure 12 we show the distance between topics. This figure is best viewed in html format (in the downloaded “output” folder) since this allows for interactive roll-overs and topic selection.

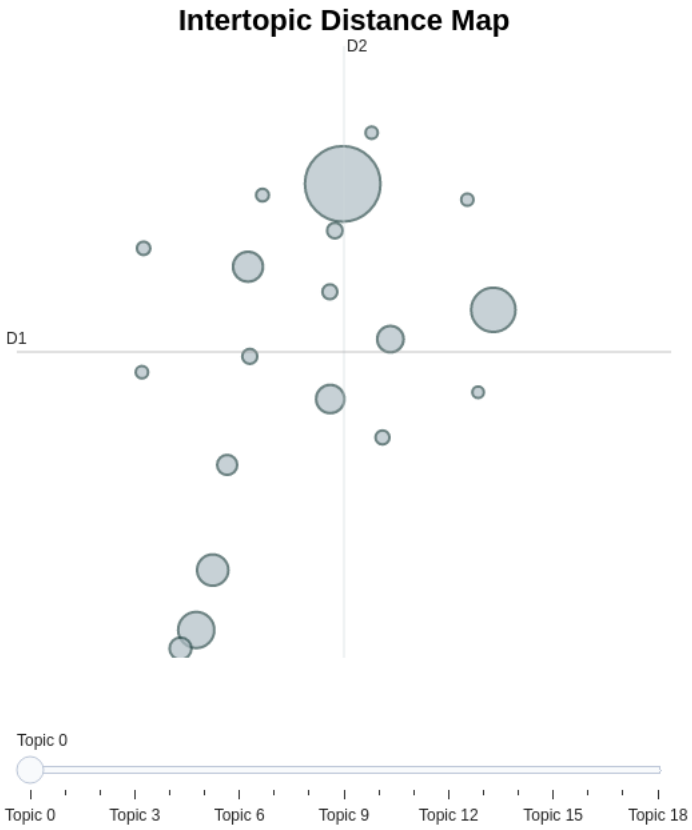


Figure 12: Inter-topic map

Figure 13 shows the relationship between topics in matrix form using a heat map. We can see, for example. 2, 3, and 4 appear to inter-related and inspection of their auto-generated labels (based on keywords) suggests there may be a sub-theme present around students.

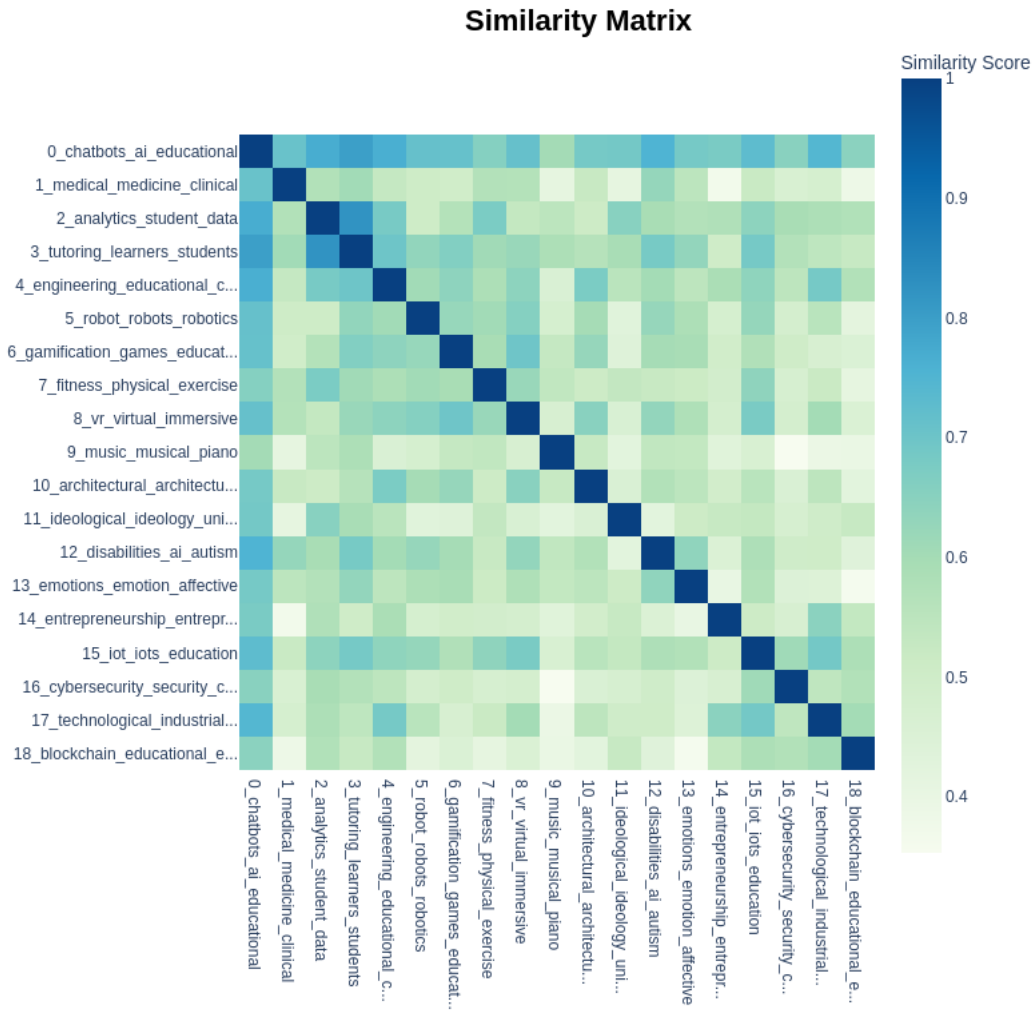


Figure 13: Topic similarity matrix

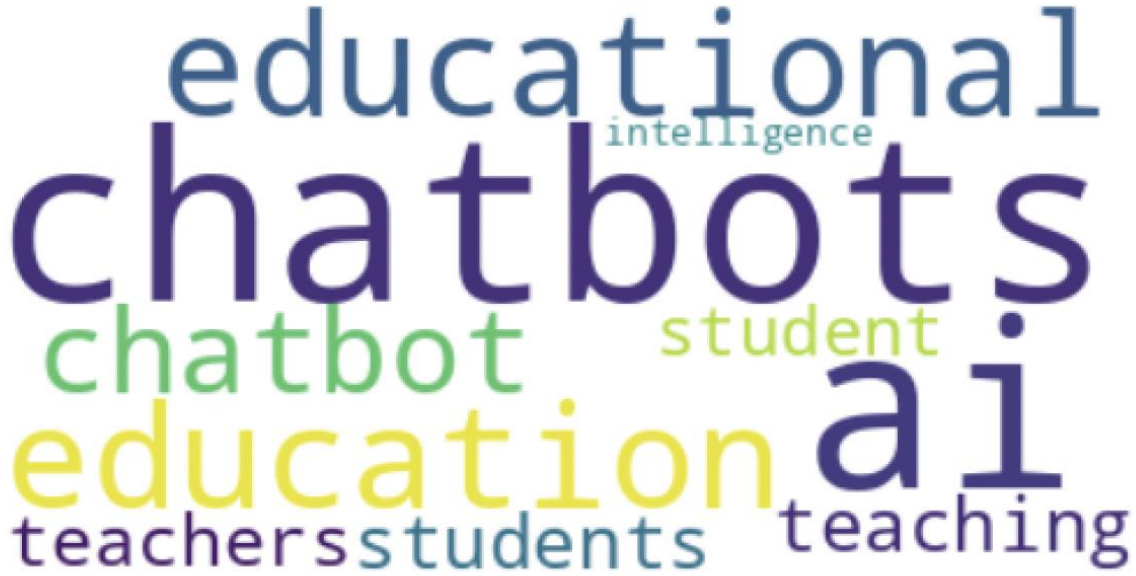
Topic labelling using ChatGPT

A pdf report of all the topics is provided to help assign topic labels (if so wished) to each of the topics (Figure 14). The report shows for each topic: word cloud, keywords, top 5 articles loading on that topic, the publication sources of the top 5 articles, and three truncated abstracts.

This information can be scanned by the researcher and suitable topic labels assigned. The information can also be loaded to ChatGPT and ChatGPT asked to assign suitable topic labels. We have found this to be very effective and a good match to human labellers (Guler et al., 2023).

Topic #0

Word Cloud



Topic Keywords
chatbots, ai, educationa, education, chatbot, teaching, students, teachers, student, intelligen

Top Sources
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
ACM International Conference Proceeding Series
Lecture Notes in Networks and Systems
Journal of Physics: Conference Series
Communications in Computer and Information Science

Top Publications
Interdisciplinary dialogue for education, collaboration, and innovation: Intelligent Biology and Medicine in and beyond 2013
Notice of Retraction: The transmutation of forming and development of Chinese traditional virtue psychology for mordern moral education
Empowering Future Teachers: Unveiling Their Attitudes and Knowledge about AI in Slovenian K-12 Education
Generative Artificial Intelligence and the Education Sector
Crowd, the Teaching Assistant: Educational Assessment Crowdsourcing

Representative Abstracts
The irruption of Artificial Intelligence (AI) based chatbot tools is undoubtedly at the frontiers of education. AI chatbots in education has emerged as a promising solution to enhance the quality of education and to improve learning outcomes. As all new technology does, it has begun to generate news about prohibition, ethical aspects, anti-plagiarism detection tools, and a series of policies from different educational systems. However, we should not deny the positive aspects of these tools if they are well used. AI-based chatbots have interesting potential to help both teachers and students, who must learn to use them well for their own benefit. This article provides an overview of AI-based chatbots, particularly ChatGPT, an artificial intelligence language model developed by OpenAI. GPT, ...
In the context of the information age, the development of artificial intelligence is flourishing, and it is becoming more and more closely integrated with economic life, which is an inevitable trend of future social development. Education is also inseparable from the support of science and technology. The penetration of artificial intelligence has changed traditional teaching models and methods. Under the influence of artificial intelligence technology, college education is developing more and more in the direction of informatization and intelligence. For college teachers, the study and application of artificial intelligence is an important and necessary means to achieve professional development. It is more important to train students to take independent self-study exams and focus on apply...
Objective: This article is dedicated to a comprehensive examination of the legal underpinnings governing the use of artificial intelligence within educational contexts. It delves into specific aspects, including chatbots, Chat GPT, issues related to chat-based plagiarism, educational simulations, and the impact of AI on the labor market. Methods: In our study "The legal framework for the use of artificial intelligence in educational activities", we use a variety of study methods to consider this issue from different approaches and take into account the key aspects that affect the effective and ethical use of artificial intelligence in education. Starting with a documentary analysis, we study the current legal norms regulating the use of artificial intelligence in the educational sphere. By...

Figure 14: Topic labelling report

Articles for in-depth literature review

By default, the CLR2 will export the 50 most relevant papers in the corpus. This is a manageable number of articles that can be reviewed by a human for the purposes of writing an in-depth literature review (we recommend the literature review contain between 30 and 50 articles).

It is also possible to override the limit of 50 and download a custom number of articles, including the entire input dataset, if required.

The CLR2 exports the relevant articles in the corpus in RIS and CSV format. By default we weight citations as 4, recency as zero, topics as 10, and individual topics are given equal weights of 5 (this can be modified by the user – see Section 6). The output in RIS format can be imported into a reference manager, such as EndNote (Figure 15), ready for in-depth analysis.

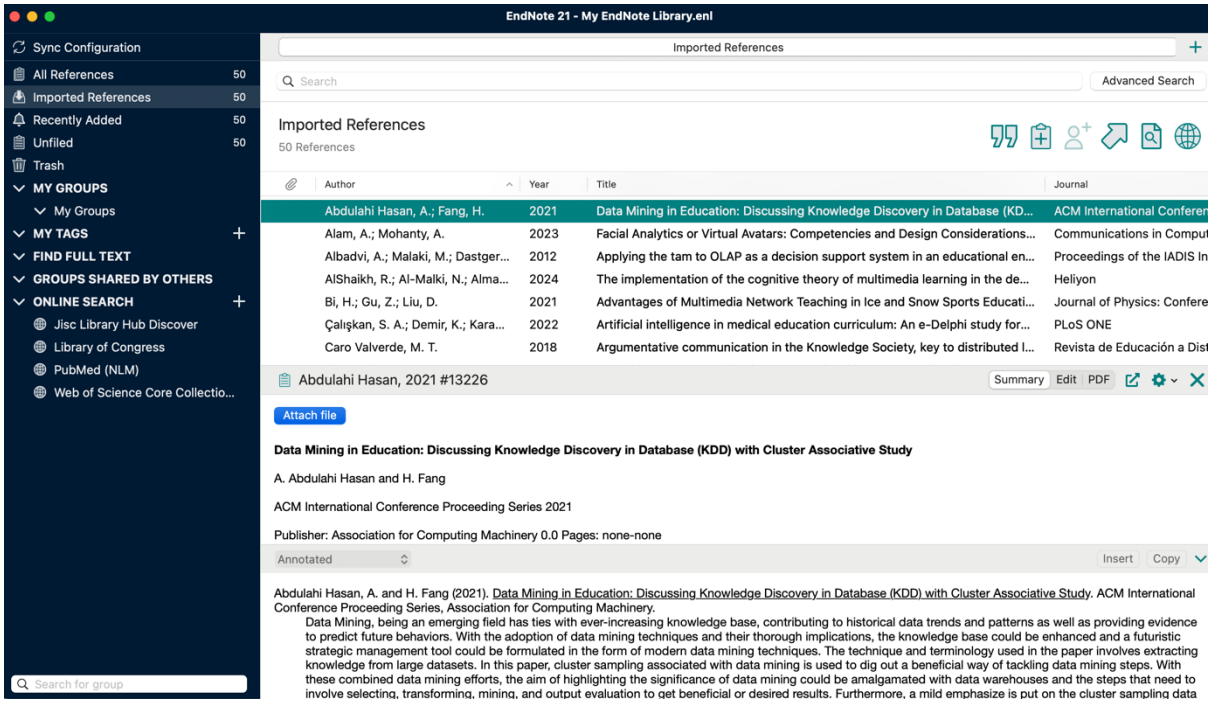


Figure 15: CLR2 output articles loaded to Endnote for further analysis

As well as providing references in the output file, the CLR2 uses custom fields in Endnote to store the results of the analysis:

C2 contains the citation score (number of citations)

C3 is a recency score (more recent articles score more highly)

C4 is the topic score

C1 is the overall score and is the sum of C2, C3, C4, providing the basis for the selection of the 50 articles to be exported from the CLR2 ready to start on an in-depth literature analysis.

6. Tuning the CLR2

We recommend running the CLR2 in standard form to see how it works and to become familiar with the outputs. You might want to tune the CLR2 to emphasise different aspects of the process. To do this, run each code segment in sequence and consider the following options.

1.1 Fast versus slow topic model

To conserve computing resources and to have a quick turnaround, the default is to run a fast topics model. If you want to apply a larger model that has been trained on academic data then run the slower model.

```
# fast topic model - run this one if you want quicker results
# if you want better (but much slower) results don't run this one and
see below.
model = topic_model(df)

# fit the model to the data
corpus, topics, probs = fit_topic_model(df, model)

# show the high level topic information
model.get_topic_info()
```

2.1 Drop topics

For the AI and education dataset, the topic model originally found 28 topics. This was reduced to 19 through the “drop topics” function.

```
# reduce the number of topics if desired
# if you want a specific number of topics then specify nr_topics = K
# (where K is the number of topics)
# otherwise the algorithm will automatically identify a suitable value
model = drop_topics(corpus, model)
model.get_topic_info()
```

3.1 Setting weights for the output articles

Global weights can be set to weight number of citations, recency of publication, and topics. By default, citations are weighted 4 and recency zero. If recency is considered to be particularly important then it might be given a weight of 9 or 10 and citations reduced to 2 or 3.

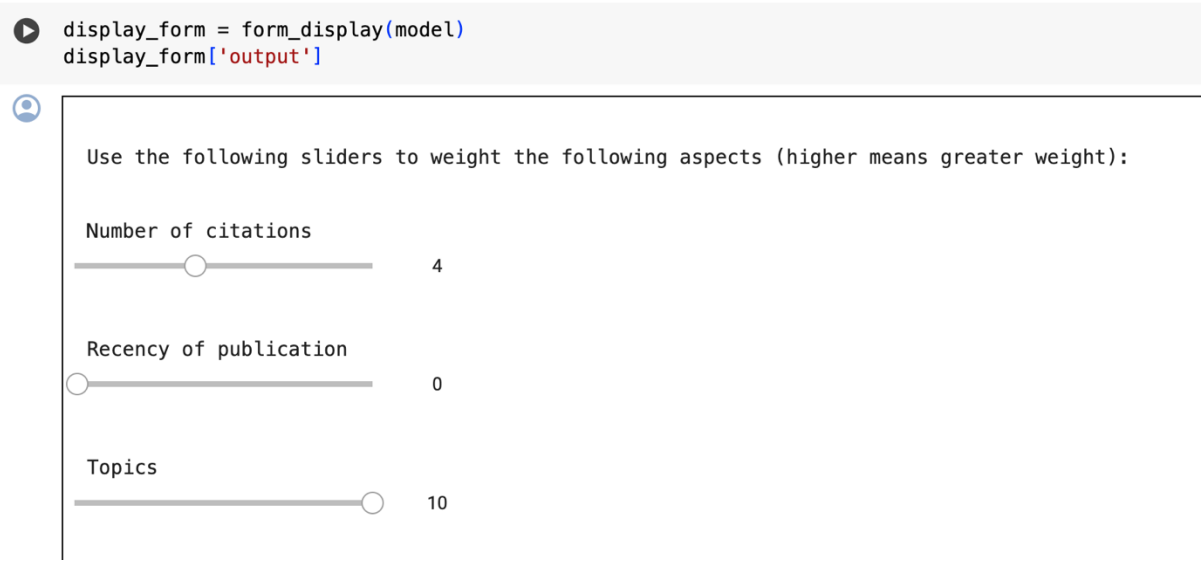


Figure 16: Setting global weights

Individual topic weights can be set (by default all topics are set to a weight of 5). Topics that are not of interest might be given a low weight (e.g., 0) and topics that are particularly relevant given a higher score (e.g., 9 or 10).

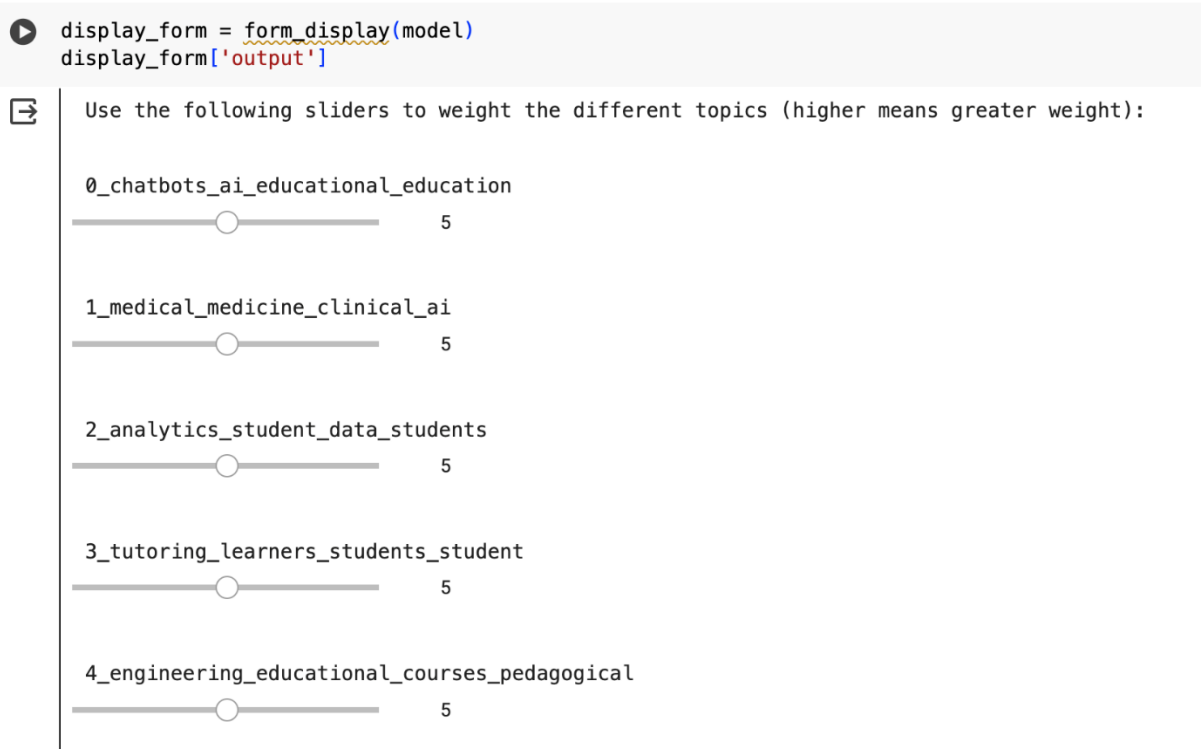


Figure 17: Setting topic weights

References

Antons et al. (2023).

Hart, C. (1998). Doing a Literature Review: Releasing the Social Science Research Imagination, Sage Publications, London.

Kitchenham, B. & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. UK.

Leeflang, P. S. & Wittink, D. R. (2000). Building models for marketing decisions: Past, present and future. International journal of research in marketing, 17, pp. 105-126.

Mortenson, M. J. & Vidgen, R. (2016). A computational literature review of the technology acceptance model. International Journal of Information Management, 36, pp. 1248-1259.

Tranfield, D., Denyer, D. & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British journal of management, 14, pp. 207-222.

White, A. & Schmidt, K. (2005). Systematic literature reviews. Complementary therapies in medicine, 13, pp. 54-60.