Mapping the South African research landscape using BERTopic

Gerrit Jo Busser
Stellenbosch University (SU)
Stellenbosch, South Africa
23637366@sun.ac.za

Marcel Dunaiski
Stellenbosch University (SU)
Stellenbosch, South Africa
marceldunaiski@sun.ac.za

Abstract

We map the South African research landscape using a dataset of scholarly articles. BERTopic, a topic model based on transformers, identifies content-based clusters of academic papers and is compared to a citation-based clustering approach that makes use of the map equation. The clusters of each approach is further analyzed with science mapping metrics.

BERTopic proves valuable as a science mapping tool in a dataset of limited citation counts. BERTopic finds meaningful clusters and identifies research themes within the dataset while the map equation approach does not. The advantages of BERTopic include its ease-of-use through c-TF-IDF scores and its adaptability to future science mapping tasks. However, BERTopic lacks mechanisms for assessing the impact of research units within the dataset. Supplementary evaluation metrics must to be incorporated for a comprehensive analysis of science using BERTopic.

Additionally, the science mapping results of both approaches are presented in a web-based interactive dashboard.

Keywords: BERTopic, science mapping, South African research, topic models

1 Introduction

When considering the state of African research, South Africa is the most proliferate [24, 31, 48]. South Africa is home to the most authors and produces the most influential publications per year [24, 48]. Analysis of South African research does exist [31]. We explore the use of two novel clustering methods to extend this analysis with a dataset of South African masters' theses and PhD dissertations.

Science mapping is a branch of bibliometric analysis that aims to observe trends in research fields. Universities, research institutes, and organizations involved in publishing academic journals require effective ways to assess research output to showcase progress and maintain a competitive edge. Funding bodies and university decision-makers seek quantitative approaches to evaluate the state of today's research [36]. The field of science mapping explores different methods that can be used for these purposes.

A topic model is a statistical model that identifies clusters of similar documents in a collection of texts. It is an unsupervised technique that is used to organize, understand, and summarize texts. Topic modelling and science mapping have an inherent link [12]. Both aim to find relationships within a text corpus that would be difficult to identify otherwise.

BERTopic [22] is a transformer-based topic model that is built on top of a pre-trained language model using BERT. BERTopic stands to be a competitive topic modeling technique when compared other conventional models [3, 22]. However, the evaluation and widespread adoption of BERTopic remains limited since it was proposed in 2022 [23].

We aim to map the science of the South African research landscape using two clustering techniques. We use BERTopic to identify content-based clusters by using the abstracts of academic research papers as input. We use the map equation [39], a more traditional science mapping technique [44], to find citation-based clusters within the same research dataset. Furthermore, we augment the analytical value of each clustering approach by using the Eigenfactor metric [5] and normalized citation indices [36]. We compare the two approaches by their ability to identify trends in South African research. An interactive web-based dashboard presents both science mapping approaches

The remainder of the paper is as follows: Section 2 provides the required background information of science mapping and BERTopic. Section 3 reviews the literature of BERTopic, science mapping and the South African research landscape. Section 4 covers the methodology of the research after which Section 5 provides both the results and a discussion thereof. Section 6 summarizes the observations into conclusions and speaks about the possible avenues for future work.

2 Background

In this section, we present background information for the main concepts used during this study. We describe science mapping and its relevant metrics. Then, we explain transformer-based language models in preparation to detail BERTopic thereafter.

2.1 Science mapping

Science mapping is a structured methodology for understanding the structure and interconnections within scientific research domains. It is the process of analyzing a corpus of text to observe trends among different units of research such as authors, journals, and research fields [13]. The datasets for this analysis often come from online platforms such as

Scopus[1] or Google Scholar[2]. Citation data is very critical for any scientific publication database as many metrics used in science mapping make use of them [10].

The goal of science mapping is to provide insight into the patterns and trends that can be observed from different metrics, often visualized with tables and graphs. Science mapping metrics are used to measure the impact of a unit of research, or to create a network relating the different units to each other.

2.1.1 The map equation. The map equation [39] is used as an agglomerative clustering method that simplifies a citation network by partitioning units of research into modules. Clusters are formed by by merging modules with similar citation patterns such that a partition that both simplifies and highlights patterns within the citation network is found. The overarching assumption is that units of research should be grouped if they cite the same sources and are cited by the same sources.

A partition M splits n nodes, or research units, into m modules. Equation 1, the map equation, gives the average description length of a random surfer's path for partition M. This equation is minimized to find the optimal partitioning of modules and is defined by:

$$\begin{aligned} L(M) =& (\sum_{i=1}^{m} q_{i\curvearrowright}) \log(\sum_{i=1}^{m} q_{i\curvearrowright}) \\ & - 2\sum_{i=1}^{m} q_{i\curvearrowright} \log q_{i\curvearrowright} - \sum_{\alpha=1}^{n} p_\alpha \log p_\alpha \\ & + \sum_{i=1}^{m} (q_{i\curvearrowright} + \sum_{\alpha\in i} p_\alpha) \log(q_{i\curvearrowright} + \sum_{\alpha\in i} p_\alpha) \end{aligned} \tag{1}$$

where $q_{i\curvearrowright}$ is the exit probability of module i, the probability that the random surfer is at any point leaving module i; and p_α is the steady state visit frequency of research unit α, the probability that the random surfer is at node α at any given point.

A deeper description as well as an optimization method for the map equation is found in Appendix A.

2.1.2 The Eigenfactor metric. The Eigenfactor score [5] is a journal influence metric that exploits the entire citation network while considering reference intensity and adding a weight of importance to citations. Reference intensity is based on the principle that a citation is more valuable from documents with short reference lists. A citation's weight of importance leads to citations from highly ranked documents having larger weights.

Although the Eigenfactor metric was developed as a journal influence metric, it can be applied to any type of research unit. The definition of the Eigenfactor metric is given in Appendix B.

[1] www.scopus.com
[2] scholar.google.com

2.1.3 Normalized citation index. The normalized citation index (NCI) measures the citation impact of a research unit. A NCI benchmarks comparative research units regardless of the publication age or field [36].

The NCI of a research unit i is defined as:

$$\mathrm{NCI}_i = \frac{c_i}{e_i} \tag{2}$$

where c_i is the number of citations received by unit i; and e_i is the expected number of citations defined as the average number of citations received by the unit type in its field during the same period.

2.2 Transformer-based language models

Many conventional topic modelling techniques, like linear discriminant analysis and non-negative matrix factorization, process textual information as a bag-of-words [22]. Transformer-based language models [45] do not have this limitation. A transformer-based language model is superior because of its potential to understand sequential elements. This does come at the cost of a reduced vocabulary size and a limited input size.

Transformer-based language models are deep learning models that embed words into vectors. A self-attention mechanism weighs the significance of each part of the input data differentially through a series of encoders and decoders. The model then captures dependencies between vectors that were far apart in the input sequence. Plainly put, transformer-based language models allow for a form of semantic analysis by embedding vectors with the context of a word as it was used in a sentence.

The language representation model BERT [16], which stands for Bidirectional Encoder Representations from Transformers, was proposed in 2018. Unlike previous models that trained left-to-right or right-to-left independently, BERT is designed to pre-train bidirectional models, which allows for a more comprehensive understanding of the context in which words appear. This bidirectional training approach enables an improved performance on a wide range of natural language tasks when compared to other models.

BERT's pre-trained models can be fine-tuned for specific tasks with relatively small amounts of data. This makes it a versatile tool for numerous natural language processing applications [43]. One of these applications is BERTopic.

2.3 BERTopic

BERTopic [22] is a neural topic modeling procedure. It leverages the power of transformer-based language models to generate topic clusters from text documents. This is done through four steps:

1. data embedding,
2. dimensionality reduction,
3. clustering, and
4. topic representation.

BERTopic takes the documents as input and organizes them into clusters that are each defined by keywords. This cluster representation yields easily interpretable topics. Another advantage of BERTopic is its modularity, allowing different approaches to be chosen for each of the steps above. Figure 1 shows the steps of BERTopic with the typical method for each; we discuss the steps thereafter.

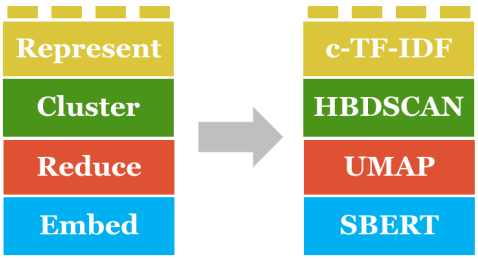

Figure 1. The building blocks of BERTopic.

2.3.1 Data embedding. The first step in BERTopic is to generate embeddings for the input documents. BERTopic breaks the texts into tokens that a pre-trained transformer model then processes. An embedding model converts a document into dense vector representations that for semantic comparison [22]. The overarching assumption is that semantically similar documents should belong to the same topic.

The sentence-BERT framework [37], or SBERT, is built on top of BERT. It is the recommended transformer model for BERTopic because it achieves state-of-the-art performance on various types of sentence and paragraph embedding tasks [38].

2.3.2 Dimensionality reduction. High-dimensional embeddings are computationally expensive to process. BERTopic reduces the dimensionality of embeddings to reduce the computational cost of sequential steps. Uniform manifold approximation and projection (UMAP) [29] is used for reducing the dimensionality of the embeddings.

UMAP is a stochastic dimensionality reduction technique that is based on the concept of manifold learning. UMAP works by approximating a lower-dimensional manifold structure using a combination of graph-based techniques and topological principles. It then optimizes this lower-dimensional structure such that local and global data characteristics are preserved. This report does not cover the extensive mathematical background required to explain UMAP.

Studies have shown UMAP to be superior to other dimensionality reduction techniques, including t-SNE and PCA, in terms of computational cost and how meaningful the reductions are [4, 25, 29].

2.3.3 Clustering. The dimension-reduced input embeddings are then clustered into groups called topics. BERTopic typically uses HDBSCAN [9], which stands for hierarchical density-based spatial clustering for applications with noise, to achieve this.

The main idea behind HDBSCAN is to group data points based on the density of datapoints. Noisy datapoints are identified as outliers and do not form part of any cluster. A high-level explanation of the HDBSCAN algorithm is presented.

1. **Build the minimum spanning tree** (MST): construct an MST on the dataset using Euclidean distance as edge weights.
2. **Build a cluster hierarchy**: sort the edges of the MST by weights in increasing order. Iterate through the sorted edges such that each edge splits the MST into two parts. Each step down the hierarchy can be seen as the potential split of a single cluster into two clusters, which we will call sub-clusters. The leaves of the hierarchy are single datapoints.
3. **Extract the clusters**: we assess the cluster hierarchy from the bottom up, using a measure of cluster stability. If a sub-cluster is more stable than its immediate supercluster, we consider the sub-cluster a potential final cluster. The size of the sub-cluster is compared to a user-defined parameter, *min_size*. The sub-cluster is considered to be a final cluster if it contains more datapoints than *min_size*. Otherwise, the sub-cluster is considered noise.

The result is a set of clustered textual documents. Each document can belong to a maximum of one cluster. Advantages of HDBSCAN include the exclusion of outliers and the use of only one intuitive parameter, *min_size*.

2.3.4 Topic representation. We would like to be able to distinguish between the different topics that were generated. BERTopic uses the documents within each cluster to model the topic representations with a modified TF-IDF measure.

TF-IDF [2] is a widely-used technique for identifying keywords in a document. A word is considered a keyword if it occurs frequently in one document but rarely across the entire corpus. Common words like articles and propositions, referred to as stop words, are likely to be filtered out. By combining two statistics, *term frequency* (TF) and *inverse document frequency* (IDF), the TF-IDF measure highlights words that can be used for characterizing a document within a corpus.

The term frequency of term t within document d is defined as

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{3}$$

where $f_{t,d}$ is number of times term t occurs in document d.

The inverse document frequency of term t in the corpus of documents D is defined as

$$\mathrm{idf}(t,D) = \log \frac{N}{1 + |d \in D : t \in d|} \tag{4}$$

where N is the number of documents in corpus D and $1 + |d \in D : t \in d|$ is the number of documents that contain term t, adjusted for division by zero.

The TF-IDF of a document can then be defined as:

$$\text{tfidf}(t, d, D) = \text{tf}_{t,d} \cdot \text{idf}(t, D) \tag{5}$$

where the most relevant keywords have the highest score.

BERTopic extends the TF-IDF concept to identify keywords in a particular topic. BERTopic concatenates the documents associated with each topic to calculate the TF-IDF on a cluster level; this metric is called c-TF-IDF (class-based TF-IDF). A topic is represented by the top n words with the highest c-TF-IDF scores of that cluster such that the contents of a cluster can easily be distinguished.

The c-TF-IDF is defined as

$$\text{c-tfidf}(t, c, C) = tf_{t,c} \cdot idf(t, C) \tag{6}$$

where each cluster c is now treated as a document and the collection of clusters C is treated as the corpus of documents.

This completes the explanation of BERTopic and all of its components. Note that the BERTopic API [21] implements all the discussed methods.

3 Related work

Topic models are reviewed with an emphasis on BERTopic. Different science mapping approaches and metrics are discussed. Furthermore, the state of South African research is surveyed.

3.1 Topic models

Topic modeling has gained significant attention as a valuable tool for discovering hidden semantic structures in large collections of text documents [7]. Srivastava and Sahami [42] mention that topic models are valuable in facilitating information retrieval and providing a high-level understanding of a corpus's content. Topic models have been applied in various domains, such as news article analysis [34] and scientific literature exploration [20]. This demonstrates their versatility in extracting insights from large-scale textual datasets. Beli [6] describes topic models as a promising component of understanding our collectively growing digitized archive.

Grootendorst released the first version of BERTopic [21] in 2020 and published its respective paper [22] in 2022. Grootendorst compared the performance of BERTopic against linear discriminant analysis (LDA), non-negative matrix factorization (NMF), Top2Vec and correlated topic models (CTM). BERTopic achieved comparatively high topic coherence scores but is consistently outperformed by CTM in terms of topic diversity. The strengths of BERTopic include its ability to scale performance with new language models and ease-of-use from the c-TF-IDF representations. However, its weakness is the assumption of each document can only belong to a single topic. Grootendorst concludes that BERTopic is a competitive topic model with stable performance across a variety of tasks.

Egger and Yu [18] find BERTopic to work exceptionally well with pre-trained embeddings. They state that its ease of use and versatility compares well against Top2Vec, NMF and LDA. The difficulty of BERTopic, a concern prevalent in all topic models, lies in the absence of objective metrics to evaluate the quality of clusters. Additionally, Egger and Yu find that BERTopic generates many outliers within the corpus. De Groot *et al.* [15] agrees that the number of outliers BERTopic generates is high. This is due to the clustering algorithm, HDBSCAN. De Groot *et al.* obtains similar results, but without outliers, when using K-means as the chosen clustering module.

BERTopic, due to its ability to use different embedding models, can process documents of different languages. Abuzayed and Al-Khalifa [1] experimented with the usage of BERTopic on a corpus of Arabic newspapers. Although they point out that BERTopic shows promise of achieving better results than LDA and NMF, they conclude that better performance measures of topic models are required. Hutama and Suhartono [23] explored how well BERTopic can identify hoax news in a low-resource language, Indonesian. The accuracy of BERTopic in the study was 0.88. Although this points to BERTopic being a competitive topic model, Hutama and Suhartono made no comparisons against other well-known topic models. Abuzayed and Al-Khalifa, as well as Hutama and Suhartono, agree that the performance of BERTopic in regard to low-resource languages should be explored further before used in practical applications.

To the best of our knowledge, BERTopic has not been extensively studied due to its relative novelty. It has, however, been used in practice. Scarpino *et al.* [41] used BERTopic and LDA to extract meaningful insights in COVID treatments in hope to find a general treatment path. They find that BERTopic demonstrates a clustering accuracy of 0.97, which outperforms the 0.92 obtained by LDA. In contrast, Ao *et al.* [3] finds LDA superior to BERTopic. They employed topic modelling to analyze how different skill requirements in job advertisements influence the wage of the job and proposed an evaluation metric based on wage regression. LDA can explain 45% of the wage variations while BERTopic can only explain 36%. The most suitable topic model seems to depend on the type of problem and the dataset.

3.2 Science mapping

The goal of science mapping is manifold and interdisciplinary.

Rosval and Bergstrom [39] summarized the 6.128 journals found in the 2004 social science edition of Journal Citation Reports using the map equation to identify the most prevalent fields of science. They summarized the journals into 88 modules using a citation network with over six million

links. Rosval and Bergstrom concluded that molecular biology, medicine, and neuroscience were the most impactful fields. Kipper *et al.* [27] identified the professional competencies required by rapidly evolving technological sectors, collectively known as industry 4.0. They did an extensive mapping of science by sorting through articles and journals from Scopus, Web of Science, and Science Direct. Kipper *et al.* analyzed the titles and abstracts of these documents to conclude that technical knowledge on data analysis, software development, sustainable development and security issues are paramount in industry 4.0. Martinez *et al.* [28] mapped the science of social work to observe how the field changed from 1930 to 2012. By using a dataset similar to Kipper *et al.*, Martinez *et al.* observes that the field of social work is increasing in publication quality and quantity. They find that modern social work research introduces LGBT, grief, and violence as major research areas that were not previously present. They observed that social work scientific discipline has become increasingly focused on the most vulnerable people in society and how people can use social services. These three studies demonstrate that science mapping can be used to summarize academic dataset and identify trends within research fields.

Chen [11] and Donthu *et al.* [17] gives a systematic review of science mapping and explains how to conduct an analysis through science mapping. The most common metrics used to map science are co-citation, co-word and bibliographic coupling. Co-citation analysis measures the frequency with which two documents are cited together, while co-word analysis is based on the frequency of co-occurrence of keywords within documents. Bibliographic coupling measures the similarity of two documents based on shared references. These metrics, along with data visualizations, are implemented by the widely used science mapping tools VOSviewer[3] and SciMAT[4].

Bergstrom *et al.* [5] disagree with the usage of quantitative measures such as co-citation, co-word, and bibliographic coupling for assessing research output. They argue that these metrics are inadequate as they disregard useful information present in full citation networks. Instead, they propose the Eigenfactor metric as a networking approach to assess research. The Eigenfactor metric takes into account the importance of each citation based on the research unit it comes from, thus providing a more comprehensive evaluation of impact. Franceschet [19] supports the use of the Eigenfactor metric, agreeing with Bergstrom *et al.* on its advantages. The Eigenfactor metric considers the weight of citations from different journals, takes into account the reference intensity of research fields, and excludes self-citations. This approach provides a more balanced and accurate assessment of research influence than traditional quantitative measures.

[3] www.vosviewer.com/

[4] https://sci2s.ugr.es/scimat/

Davis [14] compared the Eigenfactor metric against citation ranks and impact factors using a dataset of 171 medicine journals. Davis finds, with a correlation coefficient of 0.95, that the Eigenfactor metric is no different to quantitative metrics. However, West [47] contested this claim by utilizing a more extensive dataset of medicine journals. West rejected the claim that the Eigenfactor metric is equivalent to quantitative methods at the highly significant $p < 10^{-167}$ level. West raises two important points in their response. Firstly, they caution against relying solely on correlation tests to determine the superiority of one metric over another, emphasizing the need for a more comprehensive evaluation. Secondly, West highlights the importance of data visualization when mapping science, as it can reveal facets of the data that may be obscured by summary statistics.

Moed [30] advises the use of normalized citation indicators, such as NCI. Like the Eigenfactor metric, normalized citation indicators are adjusted based on different research fields and time periods. Waltman and van Eck [46] explored what the best way of normalizing citation indicators are. They used an extensive dataset of Web of Science journals to find that source-based normalization is the best way to normalize citation indicators. Source-normalizing refers to the normalization of an indicator based on its originating field or journal. Additionally, Waltman and van Eck mention that normalized citation indicators are superior to quantitative metrics like co-word and co-author.

3.3 The South African research landscape

The South African research landscape has expanded in recent years. Several studies and reports have examined the areas of research focus and the fields that are expanding the most rapidly in the country.

The South African Department of Higher Education and Training [33] evaluated the research output of universities in the country in 2019. The department reported that the number of publications increased with an average growth rate of 8.06% between 2005 and 2019. The report also details the number of publications in each field; this is summarized in Table 1.

Table 1 shows that research output in Social Sciences, Humanities, as well as in health-related sciences, makes up almost half of the total university research output of South Africa. The South African Department of Higher Education and Training commends the continuous growth of the university research outputs and states that the growth is expected to continue in following years.

Mouton *et al.* [32] provides an extensive overview of the South African research enterprise between 2000 and 2019. The number of researchers in South Africa during this time increased two-fold and the number of doctoral graduates increased by 180%. Mouton *et al.* also finds that the number of researchers per million inhabitants increased by 60%. This shows that the research capacity of South Africa is growing

Table 1. Average number of research units output per discipline for 2017 to 2019

Discipline	No. of units	% of total
Social Sciences & Humanities	4771.6	29.6
Health & Related Clinical Sciences	3070.0	19.1
Economic & Management Sciences	1527.6	9.5
Life Sciences	1731.5	10.7
Physical Sciences	1594.0	9.9
Engineering	1450.1	9.0
Agriculture	1050.7	6.5
Mathematics & ICT	866.7	5.4
Military Sciences	51.5	0.3
Total	**16112.6**	**100.0**

faster than its population. The growth in research capacity in South Africa matches the increase of total publications in the country, which was 880 in 2016. Mouton *et al.* assesses the relative field strength (RFS) index of South Africa for 2008 to 2016. This indicator compares the distribution of a country's output across fields to that of the world. This results in a single number for every field where an RFS less than 1 shows that the county's total output in that field is proportionally less than that of the world; an RFS more than 1 shows that the proportional output is more than that of the world. Mouton *et al.* presents that the South African RFS index for Agricultural Science, Social Science, and Humanities is 1.5. In contrast, the RFS index for Engineering and Applied technologies, as well as for Health Sciences, is less than 1. The RFS index of Natural Sciences is 1. Mouton *et al.* observes that South Africa research has proportionally become less focused on Agricultural Sciences and more focused on Humanities when the respective RFS indices are compared to those of 2000–2008.

Pouris and Pouris [35] conducted a bibliometric analysis of South African research output between 1990 and 2010. They found that scholarly publication trends reflect that South African universities have an increasing number of publications each year. However, universities in South Africa have a low rate of research collaborations. Similarly, Mouton *et al.* [32] observe a decline in national collaborations. Mouton *et al.* and Kahn [26] all find an increase in international collaborations, especially with the University of Oxford, the University of Cambridge and the University of London. They state that the lower rates of national collaborations lead to a limited knowledge flow and potential mobility within the South African research landscape.

4 Methodology

We discuss the approach to mapping South African research. We describe the dataset's characteristics and explain the science mapping implementation. We continue by detailing the comparison between BERTopic and the map equation as science mapping techniques.

4.1 Dataset

The dataset consisted of South African masters' theses and PhD dissertations totaling to 108.470 papers from the year 1950 to 2020. It contained the

- title,
- date of publication,
- place of publication, and
- abstract

of research papers. Each paper was assigned a document ID.

We extended our dataset with citation information by using the Microsoft Academic Knowledge Graph [5] (MAKG). We excluded papers from our dataset if their titles contained any character outside the set of uppercase letters (A-Z), lowercase letters (a-z), spaces, and numerical values (0-9). Thereafter we mapped the IDs of our dataset to the IDs used within the MAKG by matching corresponding titles. Note that this mapping was one-to-many.

We gathered citation data from the MAKG as identifier pairs. Citation data for 88.726 papers in the dataset was found. Journal and conference information was also collected for each paper in our dataset from the MAKG.

4.2 Implementation

Two pipelines were implemented to map the science of masters' theses and PhD dissertations within the landscape of South African research. A pipeline consisted of an algorithm that clusters academic research papers and supplementary evaluative metrics to analyze each cluster. We used unit tests to ensure the correctness of the implemented methods and metrics. Find more detail on the testing procedure in Appendix C.

Additionally, an interactive web-based dashboard was created to showcase the pipelines. The dashboard allowed users to review and query the dataset with respect to figures that illustrate the clustering process and the used metrics.

We used two metrics for the analysis component of each pipeline.

1. The Eigenfactor metric was used to analyze the research impact of a cluster. The Eigenfactor scores were computed on the cluster level each year with a target window of 5. Citations of papers not in our dataset or not part of any cluster were grouped into a single "other" cluster.
2. Normalized citation indices were calculated to evaluate the impact of individual research papers in the period 1950 to 2020. The citation counts were normalized per year and per cluster.

[5] https://makg.org

The two pipelines were differentiated by their respective clustering algorithms.

4.2.1 BERTopic. BERTopic was used to identify content-based clusters within the academic dataset. The abstract of an academic paper concisely summarizes its rationale and findings; we used these as input for BERTopic. We removed documents that did not have an English abstract as recommended by literature [1, 23]. The BERTopic pipeline reduced the dataset to 93.642 papers.

BERTopic was used as recommended by Grootendorst [21]. That is,

1. SBERT was used to embed the abstracts;
2. UMAP reduced the dimensionality of embeddings;
3. HDBSCAN clustered the encoded embeddings; and
4. c-TF-IDF was used to represent clusters.

HDBSCAN requires one user defined parameter, *min_size*, which specifies the minimum cluster size. We explored the resulting clusters of each $k \in \{500, 600, 700\}$.

BERTopic originally identified many stop words as cluster keywords after using c-TF-IDF. Instead of removing stop words from the abstracts, which would influence the embeddings, we removed these from the concatenated cluster abstracts to not be present during the calculation of c-TF-IDF.

4.2.2 Map equation. The map equation operates on a citation network and was used to find citation-based clusters within the dataset. Citations from papers outside of our dataset were ignored to create a closed network.

The minimization of the map equation was done through a greedy, stochastic optimization algorithm; Appendix A details this algorithm. The algorithm was applied multiple times such that the run which resulted in the best minimization of the map equation was used as the final clustering.

4.3 Comparison of clustering approaches

We performed no formalized comparison between the two approaches, as no ground-truth clusters exist that could serve as a benchmark. The focus here was largely exploratory instead, aimed at identifying trends and insightful patterns obtained by each clustering method.

Both clustering approaches are qualitatively assessed for their individual strengths and weaknesses in the context of mapping South African masters' theses and PhD dissertations. We developed an interactive web-based dashboard to allow users to directly interact with the dataset and form their own judgement on the adequateness of each clustering approach.

5 Results

We summarize the empirical findings of the BERTopic and map equation pipelines. The results of each respective pipeline is demonstrated after which the approaches are compared. The functionality and design of the interactive web-based dashboard is presented in Appendix D.

5.1 BERTopic analysis

The BERTopic procedure was applied to the dataset three times with varying values for the HDBSCAN parameter *min_size*. Each evaluation found meaningful clusters within the dataset that represent similar topics. Higher values of *min_size* lead to less clusters being formed and more datapoints being labeled as outliers. We found that *min_size* should be set to a low value to allow many clusters to be identified. Lower values of *min_size* do not diminish the quality of clusters, but rather identifies topics in academic papers that would otherwise be considered outliers. Thus, the *min_size* parameter should be set as low as the researcher is willing to analyze to maximize the depth of the study.

We expand on the results of the *min_size* = 600 run. Figure 2 illustrates the topics found and how related the contents of each cluster were. Note that the clusters in Figure 2 are not overlapping. The bubbles, which represent the size of the clusters, only overlap in the figure.

Figure 2. Intertopic distance map of BERTopic model with HDBSCAN *min_size* = 600. The dimensionality of the c-TF-IDF vectors for each cluster was reduced to two and plotted. The size of the bubbles represent the number of academic papers inside each topic. The annotated lables were subjectively assigned based on the keywords of a topic.

Each cluster identified by BERTopic is further expanded upon in Table 2. BERTopic revealed research themes that include philosophy, family psychology, and public services.

This aligned with the expectation of Social Sciences and Humanities being prevalent in the South African research landscape. Notably, education emerging as the largest topic and the mathematics not emerging as a topic did not align with literature.

Table 2. BERTopic cluster sizes and keywords. Note that the three keywords with the highest c-TF-IDF scores are shown. A better understanding of each topic was found when looking at more keywords.

Topic	No. of papers	Top 3 keywords
1	9831	study-teachers-education
2	5451	church-chapter-study
3	2331	children-study-family
4	2275	activity-extracts-cells
5	2254	project-customer-service
6	2108	development-local-government
7	2106	hiv-hivaids-sexual
8	1891	species-fish-habitat
9	1685	carbon-water-catalyst
10	1451	pain-group-muscle
11	1393	financial-market-exchange
12	1286	farmers-food-agriculture
13	1262	health-nursing-care
14	1007	plants-yield-growth
15	927	power-flow-model
16	791	conflict-international-political
17	657	rocks-formation-basin
18	605	entrepreneur-business-small

It is important to mention that each run of the BERTopic algorithm identified a substantial number of outlier research papers (≈ 50%). Although this number was large, it likely indicated that many academic papers could not be categorized into broad themes. The results are acceptable due to the identification of outliers leading to more coherent topics.

Figure 3 presents the yearly distribution of papers in each cluster. Figure 3 shows that no massive shifts in topic sizes were apparent over time. The presence of more topics could be found as time goes on.

The Eigenfactor score over topic 7 is shown in Figure 4. The scores can be interpreted as the percentage of impact on the research landscape caused by a cluster. The impact of topic 7 only starts being prevalent in the year 1983.

We can zoom into a specific cluster and identify the most impactful academic papers of that cluster using NCI. Table 3 shows the three papers with the highest NCI in topic 7.

5.2 Map equation analysis

The map equation was used to identify clusters by running a stochastic optimization algorithm numerous times. Note that the local optimum identified by the algorithm over different

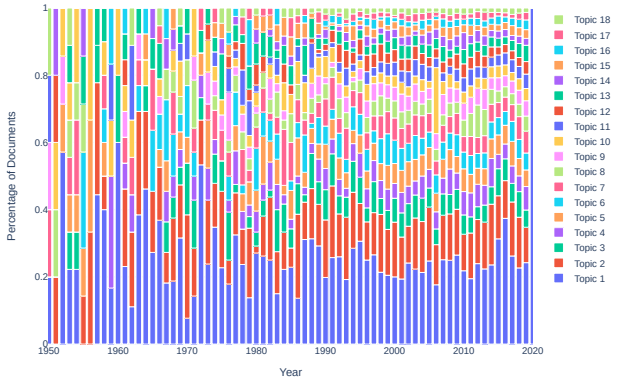

Figure 3. Document count for each BERTopic topic normalized per year.

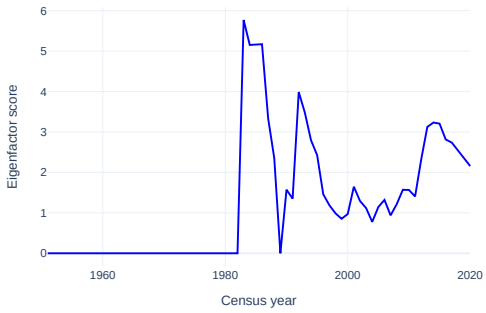

Figure 4. Eigenfactor score over time for BERTopic topic 7 (HIV/AIDS). Note that the scores were calculated with a target window of five years.

Table 3. Research papers with the highest NCI in BERTopic topic 7 (HIV/AIDS).

NCI	Year	Title
89.01	2009	The teacher as an educator within a particular culture
21.78	2005	The recognition and implementation of children's socio-economic rights in Ethiopian law
11.80	2011	Quality improvement cycle in Opuwo district hospital HIV/AIDS clinic, Kunene region, Namibia

runs did not vary much by value (less than 1% difference). The clusters found by the best run is illustrated in Figure 5.

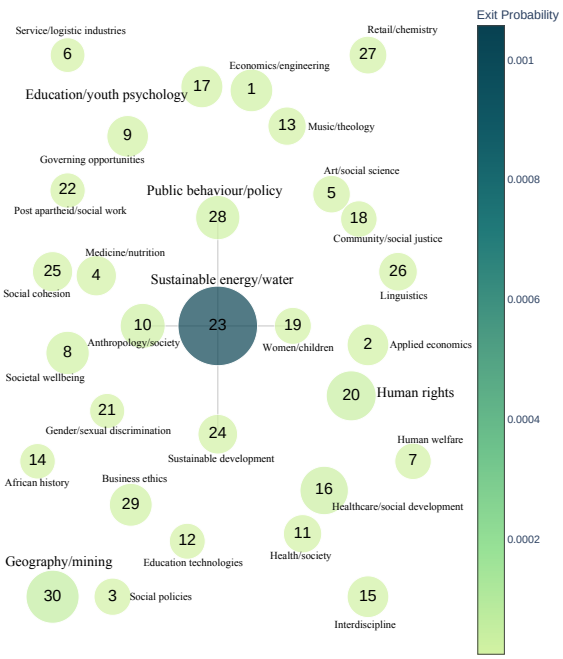

Figure 5. Citation network of South African research papers simplified with the map equation. Only nodes (clusters) containing 20 or more papers are shown. The node size is proportional to the log of its steady state visit frequency. The four edges of the graph each represent a single citation link. The annotated labels were assigned subjectively by assessing the papers within each cluster.

Table 4 further details each cluster identified by the map equation. It became clear that the map equation did not manage to find meaningful clusters. The map equation is an agglomerative clustering method. Of the 88.726 research papers for which we had citation data, the map equation optimization found 66.780 clusters. Of those clusters, 65904 contained only one research paper.

The map equation operates on citation links between academic papers. The dataset of masters' theses and doctoral dissertations contained substantially less citations in comparison to a typical science mapping dataset of academic journals. The process resulted in the identification of only a few clusters. Nevertheless, we demonstrate how the clusters were analyzed further. The impact of cluster 29 is shown in Figure 6. We look at the most impactful individual papers within that cluster in Figure 5.

5.3 Discussion

BERTopic effectively identified research themes within a corpus of academic papers. One of its distinguishing features

Table 4. The size, steady state visit probability, and exit probability of map equation clusters.

Cluster	No. of papers	Steady state(10^{-4})	Exit(10^{-4})
1	37	1.175	0.176
2	36	0.056	0.158
3	21	0.640	0.096
4	28	0.899	0.134
5	22	0.663	0.099
6	20	0.484	0.073
7	20	0.587	0.088
8	42	1.265	0.156
9	40	1.039	0.155
10	53	1.525	0.239
11	25	0.720	0.108
12	24	0.552	0.080
13	21	0.719	0.108
14	21	0.527	0.079
15	27	1.052	0.158
16	70	2.222	0.333
17	30	1.088	0.163
18	19	0.589	0.088
19	21	0.622	0.104
20	86	2.662	0.399
21	20	0.503	0.075
22	20	0.513	0.077
23	2497	72.137	1.057
24	26	0.835	0.139
25	31	0.890	0.134
26	23	0.744	0.112
27	21	0.678	0.102
28	46	1.399	0.221
29	40	1.194	0.179
30	130	3.753	0.562

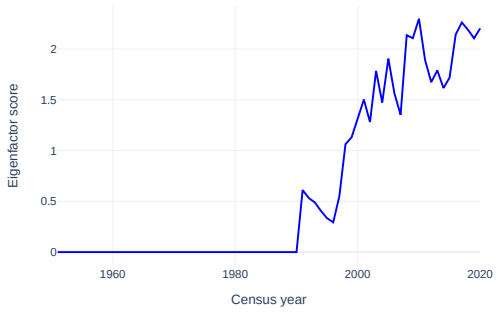

Figure 6. Eigenfactor score over time for map equation cluster 29 (business ethics). Note that the scores were calculated with a target window of five years.

Table 5. Research papers with the highest NCI in map equation cluster 29 (business ethics).

NCI	Year	Title
3.23	2004	The relationship between Total Quality Management and School Improvement
2.62	2017	Developing and testing a new generation smart meter for domestic use
2.42	2007	The impact of regulatory research structures on smaller ethics committees in South Africa

was the utility of c-TF-IDF scores, which served as a similarity measurement between the contents of clusters. The c-TF-IDF scores also present the contents of a cluster in a coherent way, which is a time-intensive and excessively subjective process when employing the map equation. These scores could potentially be applied to label clusters generated by the map equation in the future.

However, it is important to note that c-TF-IDF identifies keywords that are unique to each cluster, favoring terms that occur frequently. This bias can exclude domain-specific terms that occur infrequently within a corpus. This could account for why mathematics and other foundational sciences were underrepresented in clusters. The unique jargon of these fields would result in lower c-TF-IDF scores than terms common across general scientific jargon.

BERTopic offered notable flexibility in its application. Firstly, it accommodated varying degrees of analytical granularity through the adjustment of the *min_size* parameter. Secondly, its modular architecture suggests adaptability for different units of research. Future work could explore using different modules for embedding, dimensionality reduction, and clustering to optimize BERTopic for specific science mapping tasks.

Nevertheless, the content-based approach of BERTopic exhibited limitations. Integrating citation data to assess research impact is crucial for comprehensive science mapping. Therefore, combining BERTopic with evaluation metrics like Eigenfactor scores or NCI is imperative.

Conversely, the map equation approach was intrinsically dependent on citation data to cluster scholarly articles. This approach offered an in depth mechanism for evaluating inter-cluster impact which was unattainable with BERTopic. However, the drawback of the map equation approach was demonstrated when applied to a dataset with a sparsely connected citation network. The low citation counts of masters' theses and PhD dissertations resulted in clusters of negligible sizes and links. The map equation did not yield a meaningful summary of the dataset as a clustering method.

Our specific use case involving a dataset of South African masters' theses and doctoral dissertations demonstrated the value of BERTopic as a science mapping technique. The majority of science mapping techniques and metrics rely on citation data. Given a scenario where citations are sparse, citation-based methods like the map equation are less effective. Exploring science mapping methods that are less dependent on citation data would be beneficial to the field.

6 Conclusion

South African scholarly output occupies a prominent position in the broader African context. Identifying research trends and focal points within this landscape is essential for an understanding of its academic contributions. We evaluated BERTopic, a transformer-based topic modeling algorithm, for its efficacy as a science mapping tool applied to a dataset of South African Master's theses and doctoral dissertations.

BERTopic was employed to generate content-based clusters derived from the abstracts of academic papers. The method was compared against a citation-based clustering technique, the map equation. The study revealed the suitability of BERTopic as a science mapping approach as it effectively found meaningful research themes within the corpus of papers. The largest themes notably included education, philosophy, family psychology, and public services. Conversely, the Map Equation was less efficacious in clustering due to the dataset's low citation count.

The findings highlight the potential of BERTopic as an effective tool for science mapping, particularly when dealing with research characterized by low citation counts. However, it is imperative to conduct further analyses on the clusters generated by BERTopic, as the algorithm lacks mechanisms to assess the quality or impact of these clusters.

Future research could extend the scope of this study to include a variety of BERTopic clustering methods, each with different types of embedding, dimensionality reduction, and clustering approaches. Additionally, an opportunity exists augment the output of BERTopic to include citation data such that the impact of individual research fields on each other can be observed. Continued research in this domain could corroborate the utility of transformer-based topic models like BERTopic as viable science mapping techniques.

References

[1] A. Abuzayed and H. Al-Khalifa. 2021. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science* 189 (2021), 191–194.

[2] A. Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[3] Z. Ao, G. Horváth, C. Sheng, Y. Song, and Y. Sun. 2023. Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. *Information Processing & Management* 60, 2 (2023), 103185.

[4] R. Becht, L. McInnes, J. Healy, C. Dutertre, I. Kwok, L.G. Ng, F. Ginhoux, and E Newell. 2019. A benchmarking analysis on single-cell RNA-seq and mass cytometry data reveals the best-performing technique for

dimensionality reduction. *Nature Biotechnology* 37 (2019).
[5] C.T. Bergstrom, J.D. West, and M.A. Wiseman. 2008. The eigenfactor™ metrics. *Journal of neuroscience* 28, 45 (2008), 11433–11434.
[6] D.M. Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.
[7] D.M. Blei, Andrew Y., and M. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[8] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.
[9] R.J.G.B. Campello, D. Moulavi, and J. Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 160–172.
[10] Chaomei Chen. 2017. Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science* 2, 2 (2017), 1–40.
[11] C. Chen. 2017. Science mapping: a systematic review of the literature. *Journal of data and information science* 2, 2 (2017), 1–40.
[12] H. Chen, X. Wang, S. Pan, and F. Xiong. 2021. Identify Topic Relations in Scientific Literature Using Topic Modeling. *IEEE Transactions on Engineering Management* 68, 5 (2021), 1232–1244.
[13] M.J. Cobo, A.G. López-Herrera, E Herrera-Viedma, and F Herrera. 2011-07. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*. 62, 7 (2011-07).
[14] P.M. Davis. 2008. Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology* 59, 13 (2008), 2186–2188.
[15] M. de Groot, M. Aliannejadi, and M.R. Haas. 2022. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *arXiv preprint arXiv:2212.08459* (2022).
[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[17] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W.M. Lim. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research* 133 (2021), 285–296.
[18] R. Egger and J. Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology* 7 (2022).
[19] M. Franceschet. 2010. Ten good reasons to use the Eigenfactor™ metrics. *Information Processing & Management* 46, 5 (2010), 555–558.
[20] T Griffiths and M Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America. *vol* 101 (2004), p9.
[21] M. Grootendorst. [n. d.]. BERTopic. https://maartengr.github.io/BERTopic/index.html Accessed: May 2, 2023.
[22] M. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.
[23] L.B. Hutama and D. Suhartono. 2022. Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. *Informatica* 46, 8 (2022).
[24] M. Jeenah and A. Pouris. 2008. South African research in the context of Africa and globally. *South Africa Journal of Science* 104, 9 (2008), 351–354.
[25] D. Kabak and P. Berens. 2019. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 10 (2019).
[26] M. Kahn. 2011. A bibliometric analysis of South Africa's scientific outputs-some trends and implications. *South African Journal of Science* 107, 1 (2011), 1–6.
[27] L.M. Kipper, S. Iepsen, A.J. Dal Forno, R. Frozza, L. Furstenau, J. Agnes, and D. Cossul. 2021. Scientific mapping to identify competencies required by industry 4.0. *Technology in Society* 64 (2021), 101454.
[28] M.A. Martínez, M.J. Cobo, M. Herrera, and E. Herrera-Viedma. 2015. Analyzing the scientific evolution of social work using science mapping. *Research on social work practice* 25, 2 (2015), 257–277.
[29] L. McInnes, J. Healy, and J. Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
[30] H.F. Moed. 2009. New developments in the use of citation analysis in research evaluation. *Archivum immunologiae et therapiae experimentalis* 57 (2009), 13–18.
[31] J. Mouton, I. Basson, J. Blanckenberg, N. Boshoff, H. Prozesky, H. Redelinghuys, R. Treptow, M. van Lill, and M. van Niekerk. 2019. *The state of South African research enterprise*. DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy.
[32] J. Mouton, I. Basson, J. Blanckenberg, N. Boshoff, H. Prozesky, H. Redelinghuys, R. Treptow, M. van Lill, and M. van Niekerk. 2019. *The state of South African research enterprise*. DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy.
[33] Department of Higher Education and Training. 2021. Report on the evaluation of the 2019 universities' research output.
[34] M. Paul and R. Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. 1408–1417.
[35] A. Pouris and A. Pouris. 2009. The state of science and technology in Africa (2000–2004): A scientometric assessment. *Scientometrics* 79 (2009), 297–309.
[36] A. Purkayastha, E. Palmaro, H.J. Falk-Krzesinski, and J. Baas. 2019. Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). *Journal of Informetrics* 13 (2019).
[37] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
[38] N. Reimers and I. Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.
[39] M. Rosvall, D. Axelsson, and C.T. Bergstrom. 2009. The map equation. *The European Physical Journal Special Topics* 178, 1 (2009), 13–23.
[40] M. Rosvall and C.T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences* 105, 4 (2008), 1118–1123.
[41] I. Scarpino, C. Zucco, R. Vallelunga, F. Luzza, and M. Cannataro. 2022. Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. *BioTech* 11, 3 (2022), 41.
[42] A. N Srivastava and M. Sahami. 2009. *Text mining: Classification, clustering, and applications*. CRC press.
[43] C. Sun, X. Qiu, Y. Xu, and X. Huang. 2019. How to fine-tune bert for text classification?. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer, 194–206.
[44] N.J. Van Eck and L. Waltman. 2017. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics* 111 (2017), 1053–1070.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[46] L. Waltman and N.J. van Eck. 2013. A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics* 7, 4 (2013), 833–849.
[47] J. West, T Bergstrom, and C.T. Bergstrom. 2010. Big Macs and Eigenfactor scores: Don't let correlation coefficients fool you. *Journal of the American Society for Information Science and Technology* 61, 9 (2010), 1800–1807.
[48] N.B. Zavale and C. Schneijderberg. 2022. Mapping the field of research on African higher education: a review of 6483 publications from 1980 to 2019. *Higher Education* 83 (2022), 199–233.

A The map equation

Large systems, like citation networks, are often schematized as directed graphs [40]. It is helpful to decompose the highly complex networks such that the simplified version of the data retains important information on the network and reflect the interactions within it. Good representations both simplify and highlight the underlying structures and the relationships of the complete structure [39]. These representations are maps.

A.1 Definition

Rosvall and Bergstrom [40] describes that the problem of succinctly describing information flow boils down to a compression problem. The complexity of information flow within a network is quantified by describing a random walk through its nodes. The goal is to minimize the description length of a random walk through a network by partitioning the n nodes into m modules. A partition is denoted by M. They propose the map equation, equation 7, to calculate the average description length of a walk. The equation is given by

$$L(M) = q_{\curvearrowright} H(\mathcal{L}) + \sum_{i=1}^{m} p^{i}_{\circlearrowright} H(\mathcal{P}^{i}) \tag{7}$$

which can be expanded into

$$\begin{aligned} L(M) =& (\sum_{i=1}^{m} q_{i\curvearrowright}) \log(\sum_{i=1}^{m} q_{i\curvearrowright}) \\ & - (1+1) \sum_{i=1}^{m} q_{i\curvearrowright} \log q_{i\curvearrowright} - \sum_{\alpha=1}^{n} p_{\alpha} \log p_{\alpha} \\ & + \sum_{i=1}^{m} (q_{i\curvearrowright} + \sum_{\alpha \in i} p_{\alpha}) \log(q_{i\curvearrowright} + \sum_{\alpha \in i} p_{\alpha}) \end{aligned}$$

where $q_{i\curvearrowright}$ is the probability of going from module i to another module and p_{α} is the steady state visit frequency of node α. The steady state visit frequency of a node, p_{α}, is the probability that a random surfer will be at module α at any given point.

The surfer moves to a subsequent node β with a probability proportional to the weights of the outgoing links from node α to node β, $w_{\alpha,\beta}$, where $\sum_{\beta} w_{\alpha,\beta} = 1$ for any node α. The steady state visit frequency of a node is generated.

We start with a distribution of $p_{\alpha} = 1/n$ for each node α. The probabilities are adjusted such that the surfer has a probability $\tau = 0.15$ to "teleport" to a random node. At each iteration, we distribute a fraction $1 - \tau$ of the probability flow of the random surfer at each node α to the neighbors β proportional to the weights of the links $w_{\alpha,\beta}$ and distribute the remaining probability flow uniformly to all nodes in the network. We iterate until the sum of the differences between successive estimates of the steady state frequencies is negligible.

The exit probability of for module i can then be calculated with

$$q_{i\curvearrowright} = \tau \frac{n - n_i}{n - 1} \sum_{\alpha \in i} p_{\alpha} + (1 - \tau) \sum_{\alpha \in i} \sum_{\beta \notin i} p_{\alpha} w_{\alpha\beta} \tag{8}$$

A.2 Optimization

Given the map equation, finding an optimal node partition boils down to a computational optimization problem. It is helpful to think of nodes and modules as interchangeable concepts when minimizing the map equation. A grouped set of nodes is considered a module. However, modules can contain submodules which can each contain subsubmodules.

We explain a greedy search to optimizing the map equation as presented by Rosvall and Bergstrom [39].

1. Calculate the steady state visit frequency of each node.
2. Assign each node to a unique module and calculate the exit probabilities as described in equation 8.
3. In random sequential order, move each node to the neighboring module that results in the largest decrease in the map equation. If no movement results in a decrease, then the node remains in its original module. Repeat this procedure with a new random sequential order until no move generates a decrease in the map equation.
4. Restart the algorithm with the modules of the last level becoming the nodes of this level. This process continues until no improvements can be made.

The greedy approach can be improved. Once the algorithm is completed, each module is "reset" such that its nodes contain no submodules. The main algorithm can then be applied again with single‘node movements being allowed.

The process is stochastic and fast. It good clustering of the network can be found in a short time [39] by taking the best partition from multiple runs.

B The Eigenfactor metric

The *census year* refers to the year for which the Eigenfactor metric is calculated. The *target window* refers to the number of years prior to the census year that information is being considered. We describe the method by example of measuring journal influence as done by Franceschet [19].

Create a journal-journal citation matrix $C = (i, j)$ such that c_{ij} is the number of citations from articles in journal i in the census year to articles in journal j during the target window (generally 5 years from the census). Self-citations are ignored such that $c_{ii} = 0$. The matrix C corresponds to a weighted directed citation network in which nodes represent journals and edges represent citation links. Then, let a be the *article vector* such that a_i is the number of articles published by journal i over the target window divided by the total number of articles published by all journals over the target window.

If journal i does not cite any other journals, it is known as a dangling node. The citation matrix C is transformed into the normalized matrix $H = (i, j)$ such that all non-dangling nodes are normalized. Each element of H is defined as

$$h_{ij} = \frac{c_{ij}}{\sum_j c_{ij}} \tag{9}$$

H is then mapped to matrix $\hat{H}$ in which all rows corresponding to dangling nodes are replaced with the *article vector a*. Note that all rows in $\hat{H}$ are non-negative and sum to one. Define a new matrix P as

$$P = 0.85\hat{H} + 0.15A \tag{10}$$

where A is the *teleportation matrix* that is composed of identical rows that are all equal to the *article vector a*. The transitional matrix P can be seen as a random surfer of the journal citation network. The teleportation matrix adds a probability that the random surfer goes to a random node in the network, much like the PageRank algorithm of Google [8].

Let π be the left eigenvector of P associated with unity eigenvalue. In other words, $\pi = \pi P$. The influence vector, π, contains the scores used to weigh citation allocated in matrix $\hat{H}$. The Eigenfactor score of journal j is then

$$r_j = \sum_i \pi_i h_{i,j} \tag{11}$$

where $r = \pi H$. The Eigenfactor scores are normalized such that they sum to 100.

C Testing procedure

We created unit tests for each citation-based implementation. That is, the

1. map equation,
2. Eigenfactor metric, and
3. normalized citation indices (NCI)

were tested.

The tests served as proof-of-correctness for each of these implementations and did not cover the many adjustments and filters applied to the dataset in this study. This section presents an overview of the testing procedure without specifying each unit test. This allows a reader to verify the results of the study without detailing the individual test cases.

C.1 Test data

An arbitrary dataset was created as a test case. Table 6 and 7 present the documents and citations of the dataset respectively.

C.2 Eigenfactor metric tests

Table 8 shows the Eigenfactor scores of the test dataset. The scores align with expectation because documents of cluster 1 were cited the most, followed by those of cluster 2, and so on.

Table 6. Test data.

Document ID	Cluster ID	Year
1	1	2001
2	1	2001
3	1	2002
4	1	2003
5	2	2001
6	2	2002
7	3	2003
8	3	2004
9	4	2001
10	5	2002

Table 7. Citations matrix of test data where entry c_{rc} denotes a paper r citing paper c.

	1	2	3	4	5	6	7	8	9	10
1	0	1	1	0	1	0	0	0	1	0
2	1	0	1	0	1	0	0	0	1	0
3	1	1	0	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	1
5	1	0	0	1	0	1	0	0	0	1
6	1	0	0	0	1	0	0	0	0	1
7	0	1	0	0	0	0	0	0	0	0
8	0	0	1	0	0	1	0	0	0	0
9	0	0	1	0	0	1	0	0	0	0
10	0	0	1	0	0	0	0	0	1	0

Table 8. Eigenfactor scores of test data.

Cluster ID	Eigenfactor Score
1	30.83557
2	26.54679
3	18.561049
4	13.006324
5	11.050255

C.3 Normalized citation indices tests

Table 8 presents the NCI of the test dataset. Normalization was done by year and by cluster in alignment with the study.

C.4 Map equation tests

The map equation is optimized by running a stochastic algorithm multiple times. However, a consistent local optimum was found by running the algorithm 10 times due to the small size of the dataset.

Table 9. NCI of test data.

Document ID	NCI
1	1.2121
2	0.9696
3	1.2000
4	2.0000
5	0.9091
6	0.9000
7	0.0000
8	0.0000
9	0.9091
10	0.9000

The assigned cluster labels were ignored such that new labels were found through the map equation clustering approach. The original citation network has a random walk description length of 8.33. The map equation simplified the citation network such that this was reduced to 0.93.

Table 10 shows the cluster assignments as determined by the map equation. Table 11 and Figure 7 detail the clusters as done in the study.

Table 10. Cluster assignments of test data.

Document ID	Cluster ID
1	3
2	3
3	3
4	3
5	3
6	3
7	2
8	1
9	3
10	3

Table 11. Map equation cluster details.

Cluster	No. of papers	Steady state	Exit
1	1	0.04	0.006
2	1	0.03	0.005
3	8	0.5	0.061

D Web-based dashboard

Science mapping pipelines aim to identify trends in a research dataset. A single pipeline results in a multitude of graphs and metrics that cannot be clearly summarized on paper. An interactive web-based dashboard was implemented

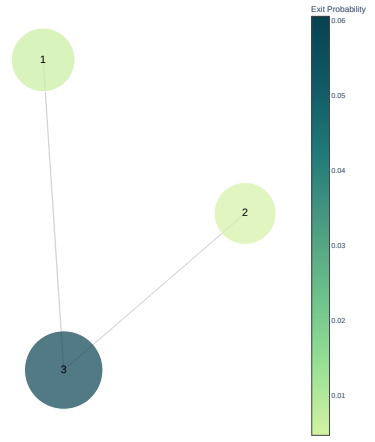

Figure 7. Test data citation network simplified with the map equation. The edge between node 1 and 3 represents two citations while the edge between node 2 and 3 represents one citation.

to showcase both the BERTopic and map equation pipelines implemented for the South African research dataset in this study. The main elements of the dashboard are presented and described in this section.

The dashboard consists of a webpage for each science mapping pipeline that a user can use to review the approach, identify trends from the graphs, and query the database. The BERTopic page of the dashboard is shown in Figure 8.

Figure 8. The BERTopic page of the dashboard. The inter-topic distance map as can be seen on the left side of the page. The right side contains a stacked bar graph representing the document counts of each topic over time; the top 15 c-TF-IDF scores and their respective keywords for the selected topic in a bar graph (which is blank as no topic is selected); and a line graph of the Eigenfactor scores over time for the selected topic (which shows the scores for the un-clustered papers as no topic is selected). A search bar and a table presenting the cluster, year, NCI, and title of the papers with the highest NCI is found at the bottom of the page.

A page consists of three or four graphs. The primary graph on the left illustrates the clusters that were found in the dataset while the subsidiary graphs detail a selected cluster.

A cluster can be selected by either clicking on the corresponding bubble in the left-most graph or by iterating through the clusters using a button. A user can hover their mouse over the graphs to see the exact values of the plotted metrics and find more detail on each cluster.

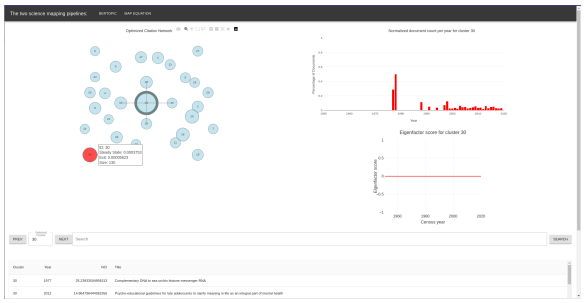

Figure 9. The map equation page of the dashboard with cluster 29 selected. The page differs to that of the BERTopic page in Figure 8 in three ways. Firstly, the left side shows the optimized citation network where the exit probabilities of a node is indicated by the thickness of its outline. Secondly, the top-right bar chart only contains the document counts over time for the selected cluster. Thirdly, no c-TF-IDF scores are plotted as this was not part of the map equation pipeline.

Figure 9 showcases the map equation page of the dashboard when a topic is selected. The paper counts over time, the Eigenfactor score over time, and the NCI table are updated to illustrate the size and impact of a specific cluster. The c-TF-IDF scores are also graphed when a topic is selected in the BERTopic page.

A user can also interact with the dashboard by searching for a term. This is seen in Figure 10. A paper is only regarded

Figure 10. The BERTopic page of the dashboard with cluster 7 selected after the term "HIV" has been searched.

if it's title contains the search query as substring. Each graph and the table is updated accordingly except for the c-TF-IDF graph. Recalculating the c-TF-IDF scores would be too computationally expensive.

The simple dashboard demonstrates the results of this study and allows users to independently assess the merits of each science mapping approach and the trends in South Africa's research landscape.