Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Kuo, Vincent; Filz, Günther H.; Leveinen, Jussi

Artificial intelligence approach for linking competences in nuclear field

Published in:
Nuclear Engineering and Technology

DOI:
10.1016/j.net.2023.10.006

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Kuo, V., Filz, G. H., & Leveinen, J. (2024). Artificial intelligence approach for linking competences in nuclear field. *Nuclear Engineering and Technology*, *56*(1), 340-356. https://doi.org/10.1016/j.net.2023.10.006

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Nuclear Engineering and Technology 56 (2024) 340–356

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

Artificial intelligence approach for linking competences in nuclear field

Vincent Kuo [a,*], Günther H. Filz [b], Jussi Leveinen [a]

[a] Department of Civil Engineering, Aalto University, Finland
[b] Department of Architecture, Aalto University, Finland

ARTICLE INFO

Keywords:
Nuclear knowledge management
Competence management
Latent semantic analysis
Community of practice
Natural language processing
Artificial intelligence
Semantic technology

ABSTRACT

Bridging traditional experts' disciplinary boundaries is important for nuclear knowledge management systems. However, expert competences are often described in unstructured texts and require substantial human effort to link related competences across disciplines. The purpose of this research is to develop and evaluate a natural language processing approach, based on Latent Semantic Analysis, to enable the automatic linking of related competences across different disciplines and communities of practice. With datasets of unstructured texts as input training data, our results show that the algorithm can readily identify nuclear domain-specific semantic links between words and concepts. We discuss how our results can be utilized to generate a quantitative network of links between competences across disciplines, thus acting as an enabler for identifying and bridging communities of practice, in nuclear and beyond.

1. Introduction

1.1. Towards more dynamic and multidisciplinary knowledge processes

Since 2009, the European Human Resource Observatory in the nuclear sector (EHRO-N) has monitored the situation with the workforce. It stresses that the competences in critical nuclear technologies are becoming difficult to sustain. According to Matselyukh et al. (2015) [1] the demography of those working in the industry, research and academia indicates that there is the danger of competences being deteriorated and ultimately lost. Furthermore, there is need for better knowledge partnerships between government, industry, education and training, science, and research communities.

Knowledge management (KM) efforts play a key role in facilitating the comparability of competences and communities. The evolution of KM since the 1980s and 1990s has shown a trend from data and information management philosophies towards more dynamic knowledge processes and collaborative innovation spaces designed to transition organizations into knowledge-based communities [2]. The current era of knowledge management is characterized by democratization and personalization of work and focuses more on heuristics, or else known as tacit knowledge. KM is viewed more and more as a social process adhering to the concepts of the Community of Practice (CoP) [3]. It is already well known that KM – involving humans, technologies and processes [4] – is by no means a discrete deterministic system, but is dynamic, fuzzy and somewhat self-organizing. It is therefore well-established that the ideal management of knowledge creation and innovation activities cannot be approached as a "factory-shop" model, where units of a discrete system are placed into boxes and assessed as such. Now, KM is characterized by a focus on stimulating factors of knowledge creation [5], and thus puts multidisciplinary linking across different communities of practice in the forefront of innovative and creative work.

The relationships between the community of practice (CoP), knowledge management (KM), and information technology (IT) have been covered extensively in literature. For instance, Von Krogh (2002) [6] investigated the role of information systems in linking CoPs, and motivated the need for a wide range of research in the respective areas of examining how IT enables communal resources through opportunities of communication, learning and knowledge sharing. Pan and Leidner (2003) [7] empirically analysed knowledge management systems in supporting the development of CoPs on a global scale, and demonstrated the importance for a knowledge intensive organization to develop a systemic capability to leverage tacit knowledge from ongoing practice and to share this knowledge within its organizational boundary. Bell et al. (2012) [8] empirically investigated internet-enabled inter-firm communication on a more strategic level, and Kietzmann et al. (2013) [9] examined how CoPs have been shaped by mobile technology.

Existing literature confirms that bridging between expert communities can be facilitated through IT based knowledge management

* Corresponding author. Aalto University, Department of Civil Engineering, Rakentajanaukio 4, 02150, Espoo, Finland.
E-mail address: vincent.kuo@aalto.fi (V. Kuo).

https://doi.org/10.1016/j.net.2023.10.006
Received 24 February 2023; Received in revised form 28 August 2023; Accepted 4 October 2023
Available online 13 October 2023
1738-5733/© 2023 Korean Nuclear Society. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

systems where people's competences are codified. However, the problem lies in codifying competences and associated communities, in such a structured way that enable semantic matching and linking. The reason is that competences are typically described in nuanced natural language, as opposed to discrete keywords or tags as an ontology. Ontological methods, involving pre-defined dictionaries and taxonomies, limit how competences can be described. Furthermore, among the diverse disciplines across the nuclear sector, similar words may be used to described different concepts, and likewise different keywords may describe similar or related concepts. Therefore, if the semantic relationships between topics are to be captured in an IT system, substantial amount of human domain knowledge is required to link the concepts, for instance, via manual tagging. A discrete keyword-based paradigm for identifying and linking competences to promote the discovery of various communities of practice is thus problematic.

1.2. State and challenges of nuclear knowledge and competence management

The International Atomic Energy Agency (IAEA) has done extensive studies on knowledge management and its implementation in nuclear organizations in practice. In particular, one investigation [10] compiles the results of numerous review visits to nuclear organizations in Member States, comprising both industry and research organizations, to conduct surveys, interviews and collect data over a 10-year period, spanning 2005–2013. In this study, the state of the adoption of various IT and knowledge management strategies have been thoroughly assessed, with focus on specific aspects such as: information management; scientific information access; tools to capture and transfer knowledge; concept mapping; collaboration tools; content management; knowledge repository; simulation tools; enterprise resource planning; portals; search engines; yellow pages; expert systems; wikis and blogs. Many shortfalls and challenges have bene identified, along with avenues for improvement in state of knowledge and competence management in nuclear organizations.

For instance, there were huge variations in how nuclear power plants (NPPs) carried out competence evaluation and management. Competency frameworks are rarely used and any formal approaches for competence mapping are mostly lacking. On the other hand, competency management appears more advanced in research and development (R&D) organizations. Competence in such organizations is also easier to measure, because the bulk of the staff is academic and is likely to regard information as public domain or open source.

In terms of knowledge management in general, related tools and techniques are mostly not integrated into the quality management system as formally written processes. It was also observed that even R&D organizations mostly do not have knowledge management processes documented and integrated within their management systems, including processes that capture learning from experience and competencies. Within the R&D organizations, there is extensive on-line access to scientific journals, citation index databases, and nuclear event information. However, there is little evidence of adoption and integration of IT solutions in support of knowledge management within NPPs. Overall, there is little evidence of alignment of knowledge management and IT strategies in NPPs and R&D organizations, and likewise little or no evidence of nuclear organizations undertaking systematic capture of tacit knowledge. Furthermore, tools such as knowledge repositories, wikis, expert systems, expert yellow pages and search engines, are rarely used. There is considerable progress yet to be made before enabling the utilization of more advanced digital approaches (e.g. concept mapping, semantic technologies, linked metadata etc.) to facilitate the knowledge and competence management processes.

Semantic technologies have permeated a multitude of domains and have naturally also been increasingly promoted within nuclear. A recent study by the IAEA in 2021 [11] focusses specifically on the long term potential of such technologies in nuclear knowledge and competence management. The report identified techniques that are particularly relevant to the nuclear domain, in order of progressiveness:

- Establishing common vocabularies, taxonomies, and thesauri.
- Integrating heterogenous knowledge sources.
- Automated indexing, categorization, and tagging.
- Semantic search and AI.
- Information and data visualizations.
- Text analytics, data mining and knowledge discovery.

The future of nuclear knowledge and competence management has huge potential for improvement with the augmentation of AI for all of the above. However, such future developments are also underpinned by the known wicked problems of data linking, interoperability, and the use of shared ontologies or taxonomies. Specifically, the manual work of establishing, updating, merging, and managing these vocabularies, metadata models, semantic structures, or other similar approaches to enable interoperability of heterogenous sources of information are time consuming and fraught with human biases. Our research addresses these challenges to encourage the utilization of more advanced digital approaches as part of nuclear knowledge and competence management solutions of the future.

2. Purpose and methodology: addressing the theory-practice gap through latent semantic analysis

It is pertinent to have effective ways to assess the transferability of knowledge and competences among many disciplines involved in nuclear. This is synonymous to mapping communities of practice across domains or between academia, research, and industry. The semantic links between competences across disciplines traditionally require a domain expert's intuition or tacit knowledge, which is a time-consuming processing and prone to inconsistencies.

In this research, we investigate how the artificial intelligence method of latent semantic analysis (LSA) can alleviate the need for human cognitive labour in linking concepts describing competences. LSA is a collection of theoretical and computational approaches that emerged in the late 1980s to early 1990s [12] as an information retrieval and natural language processing technique designed to improve library indexing and search engine query performance [13–16]. Given LSA's roots in pattern recognition and artificial intelligence, as well as psychology and cognitive sciences, it is particularly interesting to investigate as a holistic approach to competence modelling and community linking in the nuclear field.

We experiment with LSA empirically to address the theory-practice gap, as a broadly acknowledged problem in knowledge management research [17]. Thus, we also consider the situation of the real-world context of practitioners in their working environment, via preliminary interviews with nuclear domain experts. Another constraint we address is that practitioners within their organizations typically do not have extremely large, structured domain datasets at their disposal, which are typically understood to be a pre-requisite for an effective artificial intelligence application. The realms of small vs big data are not explicitly defined. Nevertheless, in this research we collect and analyse datasets of sizes that are reasonably available within organizations or institutions, as opposed to "big data" reminiscent of vast boundless data gathered from the internet or other immense repositories. By experimenting with a such datasets in our study, we demonstrate how the unsupervised LSA technique can contribute to semantic linking knowledge communities, with varying dataset sizes and big potential for scalability.

The rationale of our experimental design is as follows:

1. Preliminary study (qualitative interviews): Assess the relevance of the practical context and use-cases of competence linking in nuclear industry.

2. Small dataset analysis (mixed quantitative and qualitative): In-depth empirical exploration of the properties of LSA applied to a small dataset (N = 11) that can be read and understood manually for intuitive interpretation.
3. Large dataset analysis (mixed quantitative and qualitative): Repeat the LSA processing on a larger dataset (N = 2643) representing the potential topical landscape of nuclear knowledge and visualizing the output vectors on 2D plane for effective inspection of clusters and related concepts within specialist sub-domains of nuclear.

Our experimental design combines quantitative and qualitative methods, and is distinct in the sense that a typical data-driven algorithmic study commonly involves some benchmarking and a validation dataset to assess the outputs, while our experimental validation relates more to an augmentation of methods in qualitative case studies [18,19] often employed for knowledge management research. As such, the important aspects of our experimental logic lie in the authentic contexts in which preliminary interview results are obtained, the contents of the free form texts of both the small and larger datasets representing nuclear knowledge, as well as the choice of algorithm, that is, LSA, selected for the data processing given the distinct nature of the problem context and limitations of the data fragmentation in nuclear field. Our analysis is not primarily focussed on preparing the dataset to justify its quantitative size or coverage, since our aim is not to produce universal truths, as it would not be realistic in the real-world management of fragmented nuclear knowledge and competences. Nevertheless, we ensure that the small and large datasets analysed can demonstrate the specific properties of LSA in both cases, and that the scope and scale of the textual content would be appropriate for an equivalent methodology involving qualitative case studies, albeit utilizing LSA to process the data quantitatively instead of manually.

The subjective and tacit nature of nuclear knowledge also create challenges for using *precision* and *recall* metrics in this research, because pre-labelled validation datasets do not exist in this context for precision and recall measures to be interpreted sensibly without bias. Furthermore, our investigation is not intended on quantitatively comparing LSA to other AI-based or ontological methods in terms of the accuracy or precision of the output, but rather to empirically test the ability of LSA in processing unstructured datasets of varying sizes that represent the nature of fragmented data available in real-world settings. The validation of our experimental design is based on empirical, interpretivist assessment of whether the limited collected documentation can indeed be processed via LSA to achieve intuitive domain semantic inferences that are self-evident to the human, but traditionally challenging for computers. As such, the validation is effectively also made by the reader themselves, through observing our results and interpretations arising from semantic queries and retrievals, clustering, and visualizations presented in the discussion sections of this paper.

3. Preliminary study on practical implications

As a precursor to the algorithmic investigation, a preliminary exploratory study is done via open qualitative expert interviews with nuclear domain experts. The preliminary study aims to establish the premise, value and use-cases of competence linking with respect to the nuclear industry organizations, specifically the practitioner profiles and working environments that are relevant to competence management.

Interviewees are from the consortium of PETRUS (Project for Education, Training and Research for Underground Storage), supported by the European Commission, with the objective to promote Education and Training in geological disposal of radioactive nuclear waste. Since 2005, PETRUS has coordinated universities, radioactive waste management organizations, training providers, and research institutes to develop cooperative approach to nuclear waste disposal. The PETRUS consortium proposes strategies to ensure the continuation, renewal and improvement of professional skills by sharing resources from both academia and industries, and includes 21 representatives from 12 different countries around Europe [20]. As such, PETRUS provides ample real-world knowledge and experts in the nuclear field to validate the practical relevance of our research. Furthermore, the PETRUS agenda deals with the modelling and linking of transferable skills and competences across different organizations and sectors, thus it provides good reference point for understanding the challenges and implications of linking of communities of practice in the nuclear domain.

According to exploratory interviews with PETRUS committee members, we have identified the following practitioner profiles, and corresponding needs, addressed by the linking of competences and communities:

1) *Executive board members* need to gauge the strategically relevant competences of the organization in the long-term, as well as to constantly assess alignment of the existing combined competences of all the knowledge communities, on a high level of abstraction. These, as well as the ability to adapt quickly to new emerging competence fields, create the long-term competitive advantage for organizations.
2) *Line managers and/or technical directors* readily need to make decisions about the constituent of their own department/division/unit to ensure profitability. This implies drafting appropriate personnel from other departments as internal mobility, developing existing personnel's competences, or recruiting new human resources. The identification of knowledge communities allows technical directors to make decisions about building teams with the most complementary competences for specific new project tenders. This is challenging as teams are often multidisciplinary and located within different departments/divisions/units, and some may belong to multiple knowledge communities.
3) *Project managers* cater typically for operational efficiency within project teams and readily need to make decisions about subcontracting and/or combining existing teams with others to achieve the productivity required. Understanding the typology of knowledge communities is an enabler of sound decisions in this regard. The project manager, given the intimacy with team members, typically also conduct mentorships/apprenticeships and competence development activities. It is thus useful to identify relevant knowledge communities, in which such experiential learning and development can occur, especially in the common case that the team itself does not possess all the necessary competences to achieve certain engineering objectives.
4) *Typical technical personnel* such as engineers, scientists, researchers, technologists or general knowledge workers solve problems and complete tasks on a daily basis according to schedules and timesheets. Often, problems are complex to solve and, in addition to immediate personal social acquaintances in the workplace, there needs to be a platform where practitioners can accurately identify or discover the existing knowledge communities within the organization, from whom expert experience can be sought. Many engineering problems are repetitive and likely to have been experienced, or even solved, by other personnel at some point in the organization. Each problem solved in this manner is a lesson learnt that can be applied further and is a robust method for facilitating a self-learning organization on the operational level.
5) *Human resource* (HR) personnel cater for the human resource growth of the organization and primarily makes the recruitment decisions. HR needs to know which skills and competences are lacking or desired to be developed further in the organization, matching applicants with such skills requirements. Understanding the knowledge community constituency helps HR personnel to identify skills gaps, which in turn enables more insight regarding training and development decisions, in addition to recruitment.
6) Individuals from *professional industry associations*, unions and councils must engage in open communication with companies and understand the needs of the industry as a whole, as well as the

competence landscape embodied by communities of practice, which may change with time. The associations/unions/councils suggest standards and act as the link between industry and education/research institutions to lobby for the alignment of skills being fostered and those in demand by industry. Knowledge communities within organizations are clear indicators of the relevant professional competences in the industry.

These preliminary findings are aligned with that of knowledge management literature and show the role of communities of practice in the real-world environment. Even though the preliminary study is not central to our main research inquiry, it is significant to confirm and appreciate the valuable implications of the linking of competences in improving decision-making capacity on different levels in real-world operations for various practitioners.

4. Algorithmic steps of latent semantic analysis

The philosophies, theories, and methods of LSA are well covered in literature. For instance, Kuo (2019) [21] have been dedicated to detailed explanations and discussions of LSA, and many others have covered numerous domain applications over the years [22,23]. We describe the LSA process briefly here.

Before starting LSA, it is naturally required to gather or select an input dataset of textual data, known as the "corpus", to be analysed. In this case, the input corpus consists of textual descriptions of nuclear competences.

Generally, the LSA process can be broken down into three main functional stages (Fig. 1). The first stage consists of pre-processing, that is, parsing, filtering of texts of the input corpus into separate text passages, referred to as "documents", which consist of the words in those passages, referred to as "terms". The parsing and filtering step involves conversion of texts to the lowercase, breaking up text passages in to separate terms via tokenization, morphological stemming to remove conjugations from terms, and the removal of common high-frequency words that do not contribute semantically to the meaning of documents or passages, known as stop-words. The pre-processing stage codifies/quantifies term occurrence distributions amid the corpus of documents and constructs a term-by-document matrix (TDM), where the rows represent terms, and columns represent documents. The entries of the matrix are the occurrences of each term i in the respective document j. These entries in the term-by-document matrix, are subjected to the TF-iDF weighting transformation aimed at discounting the occurrence of frequent terms and promoting the occurrence of less frequent ones [14]. The entries of row i, column j of the TDM, are replaced by $w_{i,j}$ as follows:

$$w_{i,j} = tf_{i,j} \times idf_i \quad \text{(Equation 1)}$$

where $tf_{i,j}$ is the normalized term occurrence in each document, that is, the term frequency of term i in document j, and idf_i is the weighting factor known as the inverse document frequency, as follows:

$$idf_i = \log\left(N/n_i\right) \quad \text{(Equation 2)}$$

where N is the total number of documents in the whole corpus, and n_i is the number of documents in which term i occurs across the entire collection of documents.

The second stage, the heart of LSA, involves the application of Singular Value Decomposition (SVD), a linear algebra matrix factorization technique that is used to identify the patterns from the quantified textual representation from the first stage, that is, the weighted TDM. SVD takes any general rectangular matrix **A** with m rows and n columns and decomposes it into a product of three matrices, so that:

$$A = USV^T \quad \text{(Equation 3)}$$

where **U** (mxm) and $\mathbf{V}^T$ (nxn) are the left and right orthonormal matrices respectively. Essentially the output matrices of SVD are semantic vector representations of the terms and documents from the original explicit representation [21]. Then, dimensionality reduction can be done, whereby the length of the semantic vectors can be reduced by retaining just the most important dimensions. A very common technique for selecting how many dimensions to retain in natural language processing is to plot the squares of the singular values of the **S** matrix against the number of dimensions [24]. One then chooses the number of dimensions to keep, corresponding to where the respective squared singular values (describing the variance) decrease substantially, indicating the point where the patterns likely become insignificant. This point is also generally near the "elbow" of the graph of the squared singular values. The practical consequence of this is noise reduction. Noise reduction removes small, erratic inconsistencies inherent in term distributions and co-occurrences across the corpus. Dimensionality reduction yields truncated versions of **U** and **V** matrices, denoted as $\mathbf{U}'$ and $\mathbf{V}'$ respectively, of which the rows of these truncated matrices are the reduced semantic vectors for the terms and documents.

The third and last stage involves post-processing and further analyses. We can use cosine similarity function to measure the similarity between vectors representing terms and documents [25]. Cosine similarity is denoted by the cosine of the angle between two vectors, suppose vectors **a** and **b**, as follows:

$$\cos\theta = \frac{\mathbf{a}.\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \quad \text{(Equation 4)}$$

where **a** and **b** are two vectors of the same dimensionality, and θ is the angle between them. The cosine is acquired by dividing the dot product of **a** and **b**, by the product of their magnitudes.

This means that we can effectively quantify the semantic similarity between any two terms and/or documents. With the semantic vector space, additional testing can also be done, such as executing a query-retrieval operation, where a pseudo-document vector can be constructed based on an ad hoc input of combination of terms.

Additionally, to better inspect the semantic relationship of many LSA vectors simultaneously, it is useful to visualize the term vectors on a 2D plane, so that the Euclidean distances between them indicates their similarities. In this way, the clusters of the terms as topics in a 2D space can be easily understood intuitively. To do so, we use a popular method for exploring high-dimensional data called t-SNE (t-distributed stochastic neighbour embedding), introduced by van der Maaten and Hinton [26]. The technique has become widespread in the field of machine learning given its ability to create compelling and intuitive 2D maps from high dimensional data. The t-SNE plotting algorithms are available in the popular programming language and platform MATLAB. The semantic vectors outputted from LSA are used as input into the t-SNE algorithm, which produces an X and Y coordinate for each of the

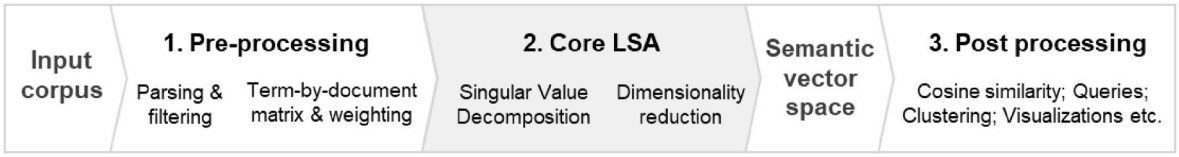


Fig. 1. General stages of the of LSA process.

vectors.

5. Small dataset collection and algorithmic analysis

5.1. Free-form texts of nuclear competences as input corpus

We first use an extremely small amount of data to demonstrate the logic of LSA in relation to the real-world problems underpinned by the data. The small dataset lends itself to test the typical assumption that artificial intelligence can only be effective given very large datasets. Addressing this has significant practical implications on the feasibility of LSA in real-world scenarios, where data are typically limited.

In this investigation, we collect 11 passages of texts of arbitrary lengths, known as "documents", representing the competences of 11 nuclear experts:

1. "The behaviour of barriers in the geological disposal of spent nuclear fuel"
2. "Nuclear emergency response planning based on decision analysis"
3. "Ageing of concrete structures in Finnish rock caverns as application facilities for nuclear waste"
4. "Solute transport modelling of geological multi-barrier disposal system"
5. "Fire simulation models for radiative heat transfer and probabilistic risk assessment"
6. "User interface for supporting operators' awareness in nuclear power plant control rooms"
7. "Systems usability concept for control room design"
8. "Interactive multi-criteria decision support - tools for practical applications"
9. "Fuel performance modelling in nuclear power plant"
10. "Code for nuclear fuel cycle analysis"
11. "Nuclear power plant procurement contracting in risky projects"

These "documents" are free form, of varying lengths, and each describes a specific aspect of nuclear knowledge, which can be specific skills or competences of certain expert functions in the nuclear field. Such textual material is available in abundance in existing organizational documentation attributing to individuals (e.g. job profiles, project descriptions, reports, correspondences etc.), as they would be in academia (e.g. course descriptions and materials, assignments etc.), and research (department descriptions, scientific articles, reports etc.).

5.2. Text parsing and filtering of competence data

We carry out the parsing using a data science platform tool called RapidMiner (https://rapidminer.com/). Since same steps can be carried out in virtually any valid textual processing tool, it is trivial which software is used. The first step is to parse the 11 documents by converting them all to lower-case and extract unique terms via tokenization. Then, we apply morphological stemming to convert the texts to the infinitive forms without inflections/conjugations. There are many existing open-source stemming algorithms used in natural language processing. One of the most widely used is the Snowball algorithm [27], which we use in our analysis. Stop-word removal is then done to filter out words such as "the", "of", "in" etc. using the standard stop-word list in the RapidMiner platform. The resulting terms, in their stemmed formats, with respect to each document, are shown in Table 1. Note, that many of the terms may appear to be strange or erroneous as if the word has been distorted or missing some parts (e.g. "emerg", "decis", "facil", "awar" etc.). These terms are indeed deliberately kept so for authentic illustration as they are the original unaltered outputs of the parsing algorithm. Here, 58 unique terms are evident in the small dataset. It is worthy to mention that since this research does not aim to investigate the differences between the algorithms used for parsing text and how it affects LSA outcomes in our case, if at all significant, the specific details and comparisons of different parsing algorithmic processes behind this step of LSA is trivial with respect to our inquiry.

Table 1
Short "documents" representing 11 expert competences in the nuclear context, and corresponding extracted terms.

Doc	Example "documents" of nuclear experts' competences	Separate "terms" parsed from the documents
Doc1	The behaviour of barriers in the geological disposal of spent nuclear fuel	behaviour, spent, barrier, dispos, geolog, fuel, nuclear
Doc2	Nuclear emergency response planning based on decision analysis	nuclear, base, emerg, plan, respons, decis, analysi
Doc3	Ageing of concrete structures in Finnish rock caverns as application facilities for nuclear waste	nuclear, age, cavern, concret, facil, finnish, rock, structur, wast, applic
Doc4	Solute transport modelling of geological multi-barrier disposal system	barrier, dispos, geolog, solut, transport, system, multi, model
Doc5	Fire simulation models for radiative heat transfer and probabilistic risk assessment	model, assess, fire, heat, probabilist, radiat, risk, simul, transfer
Doc6	User interface for supporting operators' awareness in nuclear power plant control rooms	nuclear, awar, interfac, oper, user, control, room, support, plant, power
Doc7	Systems usability concept for control room design	control, room, system, concept, design, usabl
Doc8	Interactive multi-criteria decision support - tools for practical applications	applic, criteria, decis, interact, multi, practic, support, tool
Doc9	Fuel performance modelling in nuclear power plant	fuel, model, nuclear, perform, plant, power
Doc10	Code for nuclear fuel cycle analysis	fuel, nuclear, analysi, code, cycl
Doc11	Nuclear power plant procurement contracting in risky projects	nuclear, plant, power, contract, procur, project, riski

Based on the parsed documents and terms, the term-by-document matrix (TDM) is established, consisting of the number of occurrences of a specific term in a specific document. A part of the occurrence TDM is shown in Fig. 2 with all the documents and a just few arbitrary terms for illustration.

5.3. Applying term weighting of competences using TF-iDF

The TDM thus far (Fig. 2) contains the occurrences of each term in each document, however, it could give a warped view of the sematic topical distribution within the corpus based only on the term occurrence. The reason is that the terms that occur very often in the corpus are poor to characterize the semantic feature of a single document. For instance, a term that occurs in every document of corpus does not act as a good feature descriptor of a specific document. Therefore, the occurrence of a term in a document is not directly linear to the weight of its meaning, which depends also on how frequent it occurs across the whole corpus. TF-iDF weighting (Equation (1)) addresses this by discounting the weight of terms that appear often across the corpus, as well as normalizing the occurrence of a term within one document. The latter helps to mitigate the problem that long documents could have a general advantage over shorter documents. Fig. 3 shows how TF-iDF weighting affects each term's semantic weight in one document, considering also how frequent the respective term occurs across the corpus. We use rule-based highlighting in shades of green, to better illustrate the differences.

It can be clearly seen in Fig. 3 that TF-iDF introduces nuances to the weighting. For instance, documents Doc9 and Doc10 of the occurrence TDM show that "fuel" and "nuclear" as the same weight, while the TF-iDF weighted TDM shows that "fuel" has a much higher weight than "nuclear". This makes sense, because since all the documents in our investigation are in the nuclear context, and "nuclear" appears many times across the corpus, its weight should intuitively be reduced as a feature of a single document. Similar observations can be made in other

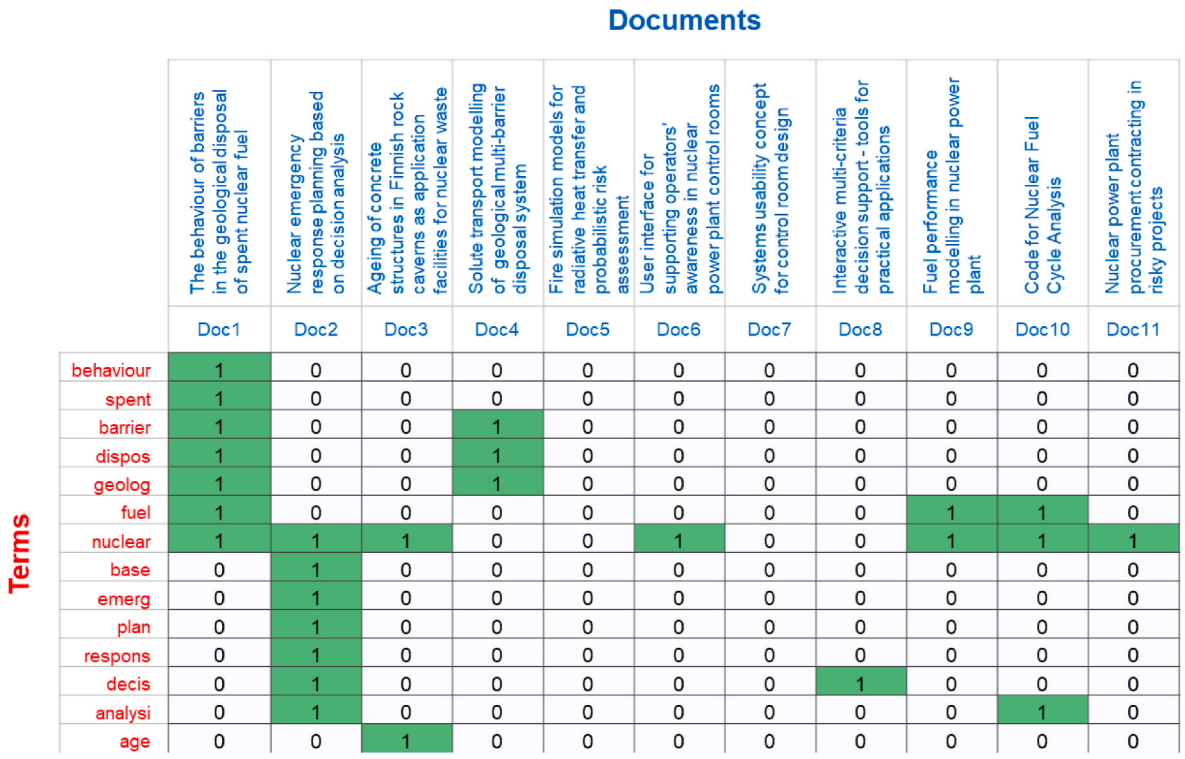


Fig. 2. Part of the occurrences TDM showing all 11 documents and a few arbitrary terms.

documents. For instance, in Doc1, the occurrence TDM indicates that all the terms in that document contribute equally to the document, (having the equal weight of 1.00), while the TF-iDF weighted TDM demotes "barrier", "dispos", "geolog", "fuel" and "nuclear" to the degree of how they are distributed in the whole corpus. The result is that more unique terms contribute more to semantically characterize a specific document. Therefore, the TF-iDF weighting in principle mimics the result of when a human underlines/highlights important words within a document, reminiscent also to the process of tagging a document with keywords. The difference is that with TF-iDF, every single term that occurs within a document would be given a weight, as opposed to a Boolean of whether a word is (or is not) a keyword. If we were to manually tag the documents in our example above, intuitively "nuclear" should not be underlined/highlighted within those documents that it appears, because the whole corpus is inherently about nuclear and therefore the term is not an important feature of the documents. This type of reasoning can occur on any level of abstraction in any corpus regarding how topics are represented and weighted.

In this small dataset, the practical output of TF-iDF weighting may not be as meaningful as in a much larger corpus of natural language text. Nevertheless, the principles are the same and the effects of TF-iDF weighting can be illustrated intuitively in our investigation. Taking this aspect into consideration can yield much more accurate semantic results down the line.

5.4. Applying singular value decomposition and dimensionality reduction

The TF-iDF weighted TDM undergoes singular value decomposition (Equation (3)) to yield the **U** (58 × 58), **S** (58 × 11) and **V** (11 × 11) matrices for this dataset. The dimensionality reduction step allows for the semantic relationships to be revealed. This is carried out by selecting to keep a certain amount of variance and discarding the rest, thus reducing the noise within the dataset. As described previously, this can be done by inspecting the singular values along the diagonal of the **S** matrix. For each dimensionality (in Table 2) we tabulate the singular values, their squared counterparts, as well as the cumulative sums and percentages. The cumulative percentages of the squared singular values thus represent the percentage of variance that is explained by the respective dimensionality, since the sum of the squares of singular values is equivalent to the total variance of the data. Dimensionality reduction reduces the variance evident in the original dataset, thereby reducing the noise, to the extent indicated by the cumulative percentage. For instance, if one chooses to reduce the dimensionality to 4 (from 11), one will keep 48 % of the variance of the original dataset.

Beside choosing the dimensionality based on the percentage variance to keep, another method is to plot the squared singular values against the dimensions, as shown in Fig. 4 for this dataset. Then, it is possible to inspect the dimensionality where the variance plot drops substantially, which usually indicates a good number of dimensions to retain.

Depending on the nature of the training data, the amount of variance represented by the number of dimensions will differ. Therefore, the shape of the graph may look different. For instance, if there is an input dataset of 1000 documents of similar domain, it is likely that there will be many more common underlying patterns than the current example of merely 11 documents. Cases involving larger corpora will typically have singular value plots, where a much higher percentage of variance is explained by few numbers of dimensions.

For this analysis, we chose to retain 8 dimensions out of the total 11, thus retaining 85 % of the variance of the original dataset (Table 2). This yields the truncated **U**' (58 × 8) and **V**' (11 × 8) matrices, where their rows represent the semantic vectors of terms and documents respectively. Fig. 5 shows a few arbitrary semantic term vectors and Fig. 6 shows all 11 semantic document vectors, all as vectors of 8 dimensions.

Therefore, SVD enables us to codify terms and documents in a dimensionality reduced semantic vector space with less noise, and at a consistent dimensionality between terms and documents. This means that cosine similarity can be used to compute the semantic similarity between any term and document with one another.

5.5. Executing query and retrieval of competences

Cosine similarity (Equation (4)) can be applied to the rows of the

		The behaviour of barriers in the geological disposal of spent nuclear fuel	Nuclear emergency response planning based on decision analysis	Ageing of concrete structures in Finnish rock caverns as application facilities for nuclear waste	Solute transport modelling of geological multi-barrier disposal system	Fire simulation models for radiative heat transfer and probabilistic risk assessment	User interface for supporting operators' awareness in nuclear power plant control rooms	Systems usability concept for control room design	Interactive multi-criteria decision support - tools for practical applications	Fuel performance modelling in nuclear power plant
		Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9
Part of the raw occurrence TDM	behaviour	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	spent	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	barrier	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	dispos	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	geolog	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	fuel	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	nuclear	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
	base	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	emerg	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	plan	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Part of the TF-iDF weighted TDM	behaviour	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	spent	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	barrier	0.11	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00
	dispos	0.11	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00
	geolog	0.11	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00
	fuel	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
	nuclear	0.03	0.03	0.02	0.00	0.00	0.02	0.00	0.00	0.03
	base	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	emerg	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	plan	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Fig. 3. Comparison between the occurrence TDM on top (same as Fig. 2), and the TF-iDF weighted TDM at the bottom. All 11 documents are shown with only a few terms for illustration.

Table 2
Singular values of **S** matrix and percentage of cumulative variance with each dimensionality.

Dimensionality	Singular values	Squared singular values	Cumulative sum	Cumulative %
1	0.382	0.146	0.146	14 %
2	0.373	0.139	0.285	26 %
3	0.341	0.116	0.401	37 %
4	0.334	0.112	0.513	48 %
5	0.330	0.109	0.622	58 %
6	0.329	0.108	0.730	68 %
7	0.308	0.095	0.825	76 %
8	0.296	0.087	0.913	85 %
9	0.254	0.065	0.977	91 %
10	0.230	0.053	1.030	95 %
11	0.221	0.049	1.079	100 %

reduced **U**′ and **V**' matrices to calculate the semantic similarities between any pair of terms and documents respectively. It is thus useful for executing semantic retrievals with respect a specified query, by ranking all the other vectors by the cosine similarity with the query.

Additionally, it is possible to make an ad hoc multiple-term query and to retrieve the most semantically related term or document. The technique entails creating a pseudo-document vector from the ad hoc query and folding it into the semantic space before carrying out cosine

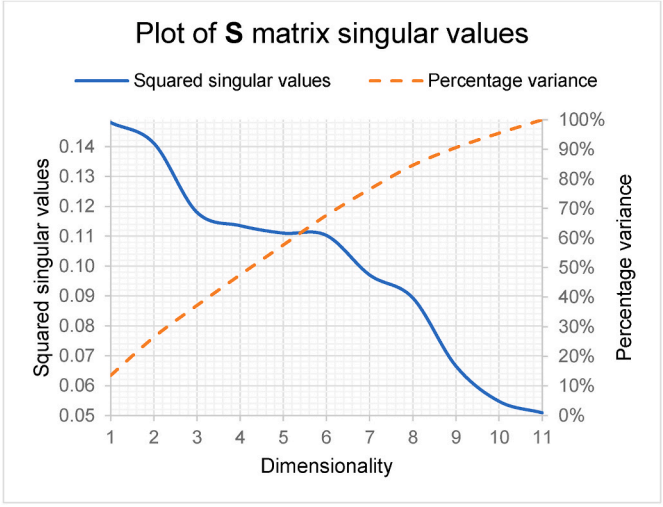


Fig. 4. Square of singular values, and cumulative percentage variance (0 %–100 %), plotted per dimensionality (1–11).

Terms	Dimensionality 1	2	3	4	5	6	7	8
behaviour	-0.089	0.058	-0.160	-0.009	-0.011	-0.236	-0.087	-0.085
spent	-0.089	0.058	-0.160	-0.009	-0.011	-0.236	-0.087	-0.085
barrier	-0.130	0.040	-0.211	-0.059	-0.018	-0.290	-0.118	-0.083
dispos	-0.130	0.040	-0.211	-0.059	-0.018	-0.290	-0.118	-0.083
geolog	-0.130	0.040	-0.211	-0.059	-0.018	-0.290	-0.118	-0.083
fuel	-0.204	0.260	-0.114	0.169	-0.017	-0.122	0.088	0.045
nuclear	-0.112	0.125	-0.036	0.034	0.074	0.016	0.018	-0.062
base	-0.096	0.176	0.161	-0.160	-0.011	0.125	-0.252	-0.177
emerg	-0.096	0.176	0.161	-0.160	-0.011	0.125	-0.252	-0.177
plan	-0.096	0.176	0.161	-0.160	-0.011	0.125	-0.252	-0.177

Fig. 5. Few examples of semantic term vectors each with 8 dimensions forming rows of truncated $\mathbf{U}'$ matrix.

Terms	Dimensionality 1	2	3	4	5	6	7	8
Doc1	-0.228	0.146	-0.366	-0.021	-0.025	-0.522	-0.180	-0.168
Doc2	-0.247	0.441	0.370	-0.360	-0.024	0.277	-0.523	-0.352
Doc3	-0.035	0.035	-0.003	-0.240	0.130	-0.008	0.667	-0.691
Doc4	-0.276	-0.006	-0.359	-0.189	-0.034	-0.436	-0.185	-0.074
Doc5	-0.064	0.027	-0.614	-0.172	-0.504	0.567	0.047	0.009
Doc6	-0.229	-0.093	-0.026	0.007	0.202	0.126	0.050	0.119
Doc7	-0.733	-0.569	0.228	0.083	-0.122	0.096	0.030	-0.027
Doc8	-0.150	0.104	0.025	-0.709	0.249	-0.033	0.256	0.545
Doc9	-0.172	0.131	-0.251	0.127	0.196	0.062	0.012	0.019
Doc10	-0.383	0.646	0.127	0.408	-0.196	-0.035	0.359	0.222
Doc11	-0.132	0.065	-0.302	0.231	0.730	0.330	-0.121	-0.060

Fig. 6. All 11 semantic document vectors each with 8 dimensions forming the rows of $\mathbf{V}'$ matrix.

similarity calculations.

Technically, a new freely defined ad hoc document vector (e.g. of a new expert competence profile) is created in the explicit space, akin to a new column of the TDM, and weighted with TF-iDF, before the vector undergoes transformation to the semantic space by multiplying by the truncated $\mathbf{U}'$ matrix and the inverse of the $\mathbf{S}'$ matrix. This enables a free open query to be made without the need to recompute the SVD. The new query, now in the semantic space with the same dimensionality, can be compared to other competences by calculating the cosine similarities to those existing vectors. This is useful if one would like to know, for instance, to which communities a new competence description, that does not exist explicitly in the corpus, may belong, or relate. As such, the cosine similarity is a quantification that also allows for clustering and/or classification.

To illustrate query and retrieval, we can take for instance the arbitrary query "nuclear risk modelling" as shown in Fig. 7. Firstly, a pseudo-document vector $\mathbf{q}$ is created and weighted using TF-iDF. It is then mapped to the semantic space as the $\mathbf{q}_s$ vector, using $\mathbf{q}_s = \mathbf{q}^T\mathbf{U}'\mathbf{S}'^{-1}$,

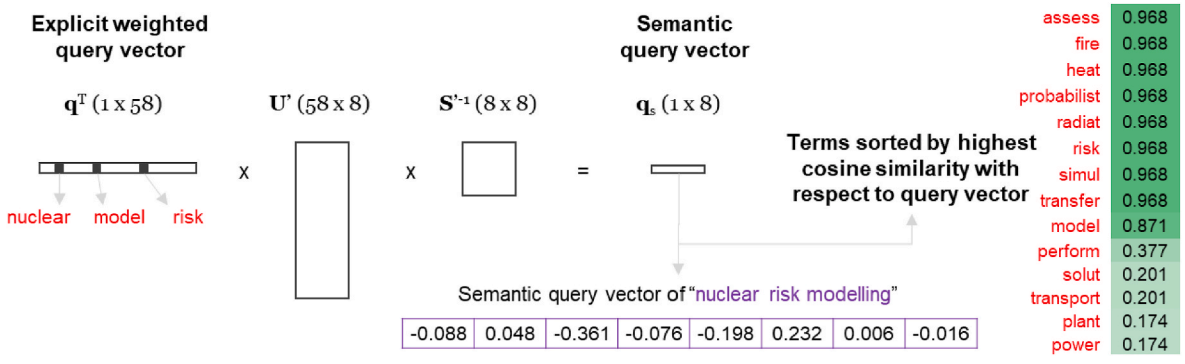


Fig. 7. Creation of the semantic query vector of the query "nuclear risk modelling".

which is simply a different form of Equation (3). Using reduced dimensionality of 8, the cosine similarities of $\mathbf{q}_s$ to all the existing terms and documents can be determined, which can then be sorted in descending order. These ranked term and document vectors are regarded as the retrieval set. In Fig. 7, the documents are filtered out of the retrieval set, showing only the terms. In many practical applications, metadata like timestamps and locations could be used for filtering.

6. Large dataset collection and algorithmic analysis

6.1. Description of the larger dataset

Much consideration has been taken to select a dataset that is broad enough to pose a realistic representation of nuclear knowledge and competences, while at the same time containing specialist vocabulary to describe specific topics in the nuclear field. It was considered to mine data from European Commission research reports on competences in nuclear, as well as IAEA technical publications regarding nuclear knowledge management. However, these were deemed inappropriate for the purposes of our research since they often present already reviewed or summarized results of larger unstructured information, and therefore, do not represent the type of fragmented corpus of documents that describes individual knowledge or competences.

It was eventually decided to use data mining to compile the larger dataset from the Nuclear Engineering and Technology (NET) Journal, given that it represents numerous facets and viewpoints across all fields for peaceful utilization of nuclear energy and radiation. Specifically, the titles of all NET articles between 2013 and 2023 are mined and compiled to form a corpus of 2643 documents for LSA. The full input dataset is available via Mendeley Data (https://data.mendeley.com/datasets/9j6std925r/1) [28].

6.2. Latent semantic analysis and dimensionality reduction of large dataset

The LSA steps will only be described briefly in this section as the process is identical to the process described in Sections 4. and the empirical testing of the small dataset presented in Sections 5.2. The 2643 documents are parsed in the same way to obtain 3653 unique terms, thus creating the respective term-by-document matrix of 3653 × 2643 to be transformed via TF-iDF weighting. The TF-iDF weighted TDM then undergoes singular value decomposition to yield the **U** (3653 × 3653), **S** (3653 × 2643) and **V** (2643 × 2643) matrices for this dataset.

The squares of the singular values of the **S** matrix are plotted against the dimensionality to reveal a clear "elbow" of the graph (Fig. 8), that is, the apex of the curve where the dimensionality decreased sharply and starts to flatten out. The corresponding dimensionality at that point (100 dimensions) has been chosen to be retained, thus yielding the truncated **U'** (3653 × 100) and **V'** (2643 × 100) matrices, where their rows represent the semantic vectors of terms and documents respectively.

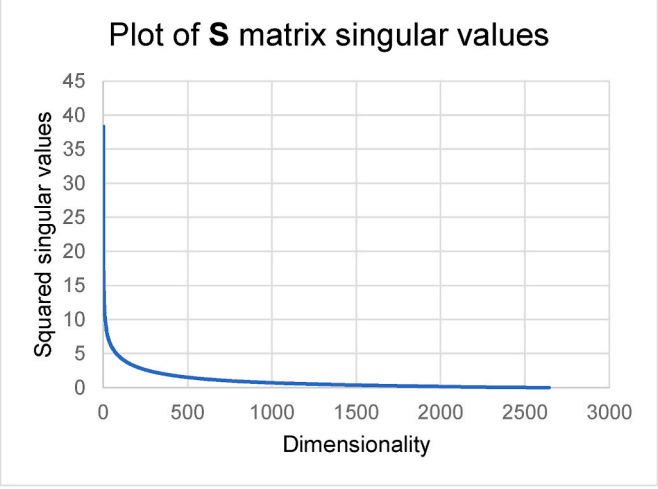


Fig. 8. Square of singular values plotted per dimensionality (1–2643).

The truncated **U'** matrix with all the term vectors of 100 dimensions, and the diagonal singular values of the **S** matrix and squared counterparts, are available via Mendeley Data (https://data.mendeley.com/datasets/9j6std925r/1) [28].

6.3. Preparing top 500 vectors for visualization in 2D using t-SNE

The semantic vectors are now available for inspection, to qualitatively assess the intuitiveness of their relationships based on nuclear domain. For practicality and ease of inspection, we have ranked the top 500 terms (of the total 3653) by the summed TD-iDF of each across the corpus, roughly representing the most prevalence or important topics in the topical landscape of the corpus (more in Section 7.3). The 500 terms, each of 100 dimensions, are used as input to the t-SNE algorithm, which we can use to obtain a reduced 2D plot of the vectors, so that their Euclidean distances to one another roughly represents their semantic relatedness visually. It is noted that there would be information loss when plotting a high dimensional vector space onto 2 dimensions of X and Y coordinates, thus the visualization is considered an approximation. Nevertheless, the ability to visualize larger number of vectors as points in a 2D plane makes it possible to inspect for clusters and relationships very easily, to inform the nature of the inferred semantics within these vectors with respect to the nuclear domain. Section 7.4discusses the interpretations of these results in more detail. Furthermore, the X and Y coordinates of all 500 terms as outputs of t-SNE, as well as their 2D visualisations, are available via Mendeley Data (https://data.mendeley.com/datasets/9j6std925r/1) [28].

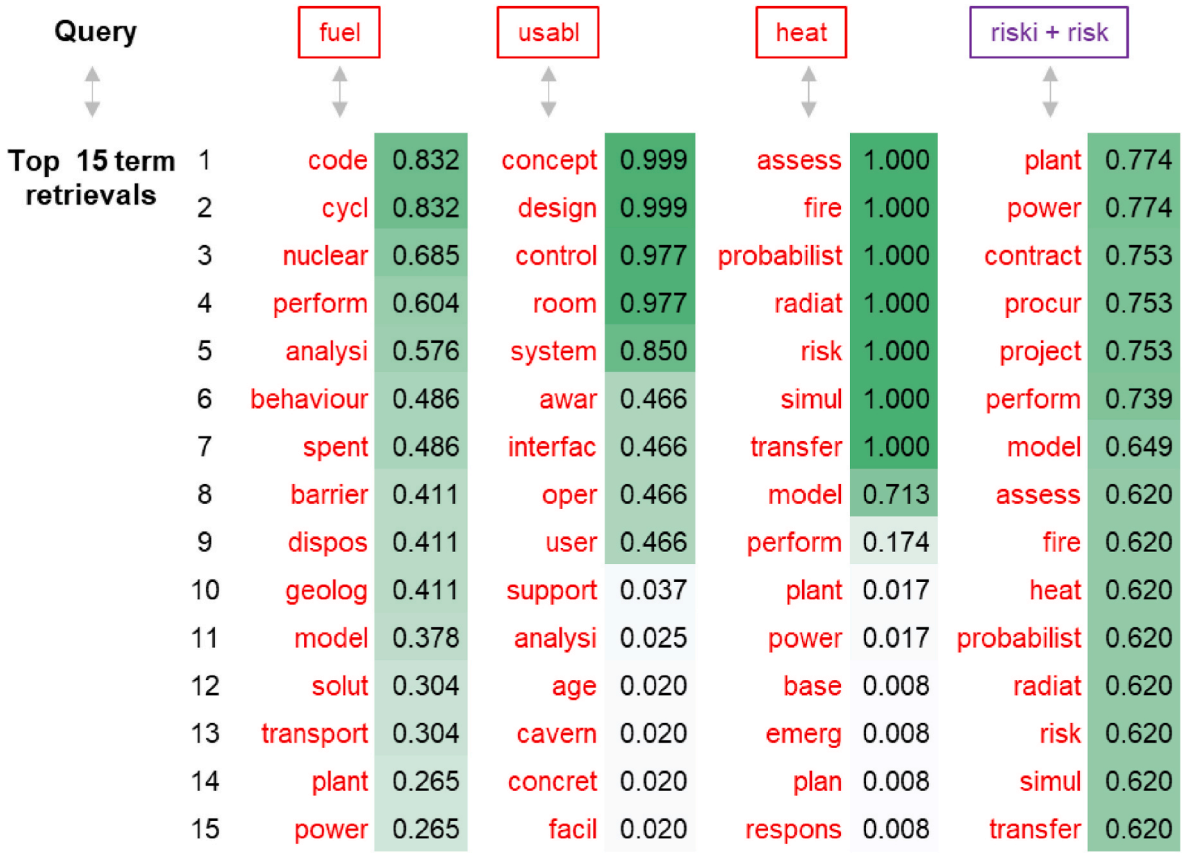
7. Results and discussions

7.1. Interpreting term-term cosine similarity of small nuclear competences dataset

Semantic vectors, presenting terms and/or documents, can be compared to one another using cosine similarity, as a measure of their semantic distance. This capability of LSA addresses the computational and automatic handling of competences in natural language textual descriptions, as an indicator of the respective knowledge communities. In practice, an obvious use-case of such capability is semantic information retrieval, as query-retrieval cycles. We discuss the query-retrieval mechanics in depth by looking at a few example queries and their corresponding results, to give an idea of the nature of the semantic inference and behaviour of the algorithm.

Taking a few arbitrary query terms: "fuel", "usability", "heat", and "risky", each query and respective term retrievals are shown in Fig. 9. The retrievals are simply terms with the highest cosine similarities to each query at a dimensionality of 8, thus retaining 85 % of the original variance. Note that the query terms shown in Fig. 9 are the stemmed versions of the query, therefore the query "usability" is stemmed to "usabl", and the query "risky" is stemmed to both "riski" and "risk". The single term queries (i.e. "fuel", "usability", "heat") are represented simply by the respective semantic term vectors, while the "riski + risk" query vector is generated from a new pseudo-document vector of "riski + risk" in the semantic space (described in Section 5.5). For each query, the top 15 retrievals are tabulated, based on sorting the terms by the cosine similarity values with respect to the query in descending order.

Even for non-experts the semantic relationships between "fuel" and retrieved semantic terms are self-evident. Concepts of "fuel performance code", "fuel cycle", "nuclear fuel", "spent fuel" etc. are well-known. The output has also picked up the significance of "fuel behaviour" and "analyses" thereof.



	fuel		usabl		heat		riski + risk	
1	code	0.832	concept	0.999	assess	1.000	plant	0.774
2	cycl	0.832	design	0.999	fire	1.000	power	0.774
3	nuclear	0.685	control	0.977	probabilist	1.000	contract	0.753
4	perform	0.604	room	0.977	radiat	1.000	procur	0.753
5	analysi	0.576	system	0.850	risk	1.000	project	0.753
6	behaviour	0.486	awar	0.466	simul	1.000	perform	0.739
7	spent	0.486	interfac	0.466	transfer	1.000	model	0.649
8	barrier	0.411	oper	0.466	model	0.713	assess	0.620
9	dispos	0.411	user	0.466	perform	0.174	fire	0.620
10	geolog	0.411	support	0.037	plant	0.017	heat	0.620
11	model	0.378	analysi	0.025	power	0.017	probabilist	0.620
12	solut	0.304	age	0.020	base	0.008	radiat	0.620
13	transport	0.304	cavern	0.020	emerg	0.008	risk	0.620
14	plant	0.265	concret	0.020	plan	0.008	simul	0.620
15	power	0.265	facil	0.020	respons	0.008	transfer	0.620

Fig. 9. A few query inputs with corresponding returned semantically related terms. All the terms are displayed as their stemmed versions, that is, their tokenized forms used in the analysis.

"Usability" retrievals are also obvious, that is, linked with "concept" as an attribute of the "control room" and "systems". "Usability" is a consequence of good "design" and thus are semantically related. We also see that lower down, the connotations of "interface", "operation", and "user" are inferred, even though those are syntactically different to "usability", it is intuitive that they are conceptually linked and has been identified quantitatively.

Relation between "heat", "fire", "radiation" and other highly ranked cosine entities are obvious, such as the connotation of "heat" as a "risk" and heat "transfer" as a subject of "simulation". The "probabilistic", "assessment", and "modelling" terms indicate the context for the "heat" topic as it appears in the corpus.

Combined "risky + risk" are related to "contracts" and "procurement" contexts within "power plants"; as well as "heat" and "fire". It is interesting to observe that the term "fire" is returned for both "heat" and "risky" queries (with cosines of 1.000 and 0.620 respectively). Intuitively, fire is indeed semantically related to heat and risk. Furthermore, it is noteworthy to see that "risk" has been characterized with multiple connotations, made possible by both the stemming and LSA. Thus, according to our small demonstration corpus, "risk" query has strong "contractual" dimensions, while also having a physics dimension through the "fire" retrieval.

Looking at the way cosines decrease for each query (gradual brightening of the green backgrounds in Fig. 9) also indicates how the topic is distributed across the corpus. For instance, it is apparent that "fuel" and "riski + risk" are distributed broader semantically, showing more gradual cosine decrease, than "usability" and "heat", of which the cosines drop more substantially at one point. This is an indication of ambiguity of terms inherent in the corpus. With our small corpus of the 11 documents, the LSA recognizes that "fuel" and "risky" has more connotations, that is, more related topics or contexts in which they can be interpreted, while "usability" and "heat" are both more specific and thus independent topics in the corpus. The pattern of the decreasing cosine can also be interpreted as a measure of the semantic centrality and density of certain queries. The more gradual the cosine reduces relatively, the broader or more ambiguous the query. Therefore, the more connotations a query has, the more gradual the decreasing cosine values would appear. This is the capability to capture fuzziness of how concepts are linked amid natural language description of nuclear competences, expressed quantitatively.

This makes it possible for LSA to seemingly link terms and documents as if the meanings of the texts are considered. Even syntactically unrelated terms can be linked if they have similar meaning, that is, based on how they are distributed with respect to all the other terms and documents in the corpus through co-occurrences over any number of intermediate links. The algorithm captures consistent terms' co-occurrences over any degree of separation across documents. This would be difficult for a human to identify when terms are co-occurring with many degrees of separation, albeit consistently.

7.2. Interpreting term-document cosine similarity of small nuclear competences dataset

To enrich the understanding of queries, it is pertinent to observe the term-term retrievals alongside term-document retrievals. This helps to illustrate some deeper intuitive logic on how the retrievals work.

In Fig. 10, the top 5 similar documents for each query are shown, along with the explicit occurrences of the query terms coloured in red within each document for ease of reference. Naturally the top documents for each query will usually be those that explicitly contain the term/s of the query. This makes sense, since if a document contains a specific term, it is assumed to embody a semantic concept related to that term. It

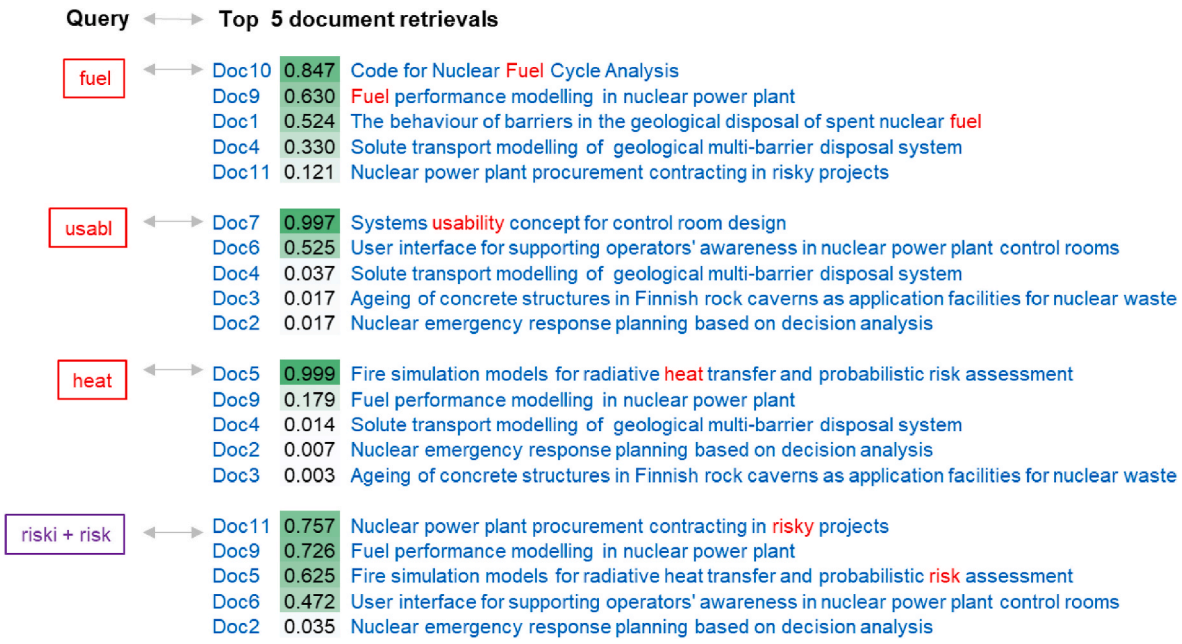


Fig. 10. A few queries with top 5 semantically related documents ranked by their cosine similarity values. Explicit occurrences of the query term/s within the document are coloured red for ease of inspection.

is also easy to notice that the top retrieved terms from Fig. 9 usually are found within the same documents together with the query term, as shown in Fig. 10. Therefore, queries and respective top retrieved terms typically co-occur in documents of the corpus.

We can interpret Fig. 10 by looking at each query and the corresponding top retrieved documents. Taking, for instance, the "fuel" query, Doc10, Doc9, and Doc1 all contain the term "fuel". However, Doc4 and Doc11 do not, but reflects some, albeit little, correlation to "fuel" based on the co-occurrence of other terms. For instance, Doc4 shares "modelling" and "geological" with Doc9 and Doc1 respectively, while Doc11 shares "nuclear power plant" with Doc9. One way to verbalize this interpretation is that Doc4 is semantically related to fuel, because there is modelling involved, although about solute transport, and performance modelling of fuel is a strong concept within the corpus. Furthermore, since disposal of nuclear fuel (nuclear waste) is in geological systems within this corpus, the fact that Doc4 relates to geological disposal context also implies its relation to fuel. Of course, there could be other, less obvious, semantic links that adds to the tendency for Doc4 to be retrieved with "fuel". Nevertheless, this interpretation is helpful to intuitively understand the principal rationale behind the retrievals based on LSA.

The same reasoning is used to interpret the documents returned for other queries. For instance, "usability" appears explicitly in Doc7, but Doc6 is also ranked high despite it not containing "usability", primarily because of the "control room" topic, which it shares with Doc7.

For the query "heat" the 2nd ranked Doc9 does not contain "heat" but has "modelling" which also occurs in the highest ranked Doc5 containing. The reason that Doc9 is highly ranked can be interpreted to be due to the shared topic, so that heat simulation and fuel performance are linked based on them being both modelling techniques.

For the query "risky" ("riski + risk"), same logic is observed there, but it is worthy to note that even though Doc5 contains "risk" explicitly, Doc9 (without "risk"/"risky") is still ranked higher than Doc5, albeit not by much. This is due to the occurrence of "nuclear power plant" in Doc9, which also occurs in the highest ranked Doc11. This shows the reasoning capability of LSA, to codify the relevance of retrievals via cosine similarity through the co-occurrence of other related concepts in documents of the corpus, rather than just the explicit occurrence of query terms. This feature is also the clear differentiator to keyword-based retrieval techniques, which do not consider the semantic distribution of bodies of text, but handles ranking based on texts' explicit occurrences.

With the small dataset in this analysis, it is easy to observe the behaviour of the LSA technique. With larger datasets, even though the same mechanics apply, it would be more difficult to explain visually and the layers between input and output of the query-retrieval step may become a "black box". In the case of the latter, however, it is still possible to infer the semantic correlations between terms and or documents intuitively, but sometimes, one may not know exactly the reasoning or rationale at first glance and may then require more exploration for the human to gain understanding. This contributes to the capability of LSA as a recommendation system, by which the user, even if not provided the answer, is assisted in formulating questions given hints where it may be interesting to probe.

Looking at all the patterns of decreasing cosines of the queries and retrieved documents in Fig. 10, visualized in shades of green, it is also clearly apparent that the term-term and term-document retrieval results are related. Generally, more documents in the corpus have relevance to "fuel" and "risky" as a topic, than "usability" or "heat". "Fuel" and "risky" retrieves more documents from the corpus, but with lower specificity, given the lower and gradually distributed cosine values of the retrieved entities. On the other hand, "usability" and "heat" get very few hits, but with very high specificity, that is, high cosine scores. In this manner, queries can be tabulated in the same fashion to compare the semantic distribution of documents with respect to any query. For instance, it is possible to see whether a specific query topic, represented by a term or several terms, is more widespread and shared by many documents, or otherwise only covered by few documents. This, along with the cosine similarity values and colour visualizations, makes it simple to quickly gauge the semantic relevance of documents with respect to a query, in addition, also to understand the nature of the topical distributions. Such is the nature of semantic summarization that would typically require domain-specific human reasoning to achieve – to read, understand, arrange, and consolidate the texts. We can now use LSAto distil such insights to aid decision-making and support a variety of information retrieval applications underpinning sense-making and knowledge management.

The current LSA's "intelligence" is limited to that of the 11 documents of the nuclear competences, as if that is all that the system

"knows". Retrieval results may give unintuitive answers if the training dataset does not cover the extent of the knowledge required for the query. For instance, if a query is made about the financial domain, using this small nuclear competence dataset to infer the semantic structure, the output would likely not return meaningful results since the dataset does not embody financial knowledge, and thus doesn't exhibit the semantic structures applicable in the financial domain. Naturally, the query potential of the system is limited by the scope of the input training dataset, which must thus be sufficient with respect to the needs of the query use-cases. Despite the apparent limitation, this is also considered the virtue of LSA to capture contextuality. It is fundamental in searching and retrieving relevant knowledge within the subjective nature of knowledge domains, which has been a principal challenge in knowledge management software systems. In terms of handling nuclear competences and attempting to computationally capture semantics, these query-retrieval interpretations give the basis upon which further analyses can be done towards mapping of communities of practice, to the extent that would have required human tacit knowledge to carry out.

7.3. Top TF-iDF terms as topics of the large dataset

Furthering the discussion around the small 11 document dataset, we now turn our attention to the larger dataset consisting of 2643 documents described in Section 6. While the small dataset can enable in depth inspection and interpretation of the query and retrievals, the larger dataset is a better representation of the variety of topics in the nuclear field on a realistic scale.

After parsing the 2643 documents into 3653 unique terms and weighting them using TF-iDF to adjust for the effects of common vs unique terms, it is possible to estimate the most important topics simply by summing the TF-iDF values. For example, the top 20 terms ranked by the TF-iDF are as follows in descending order: "nuclear"; "reactor"; "analysi"; "use"; "fuel"; "system"; "power"; "base"; "model"; "studi"; "neutron"; "plant"; "effect"; "method"; "develop"; "simul"; "design"; "thermal"; "water"; "evalu"; and so forth. Of course, these form a very reductive overview, but are clearly intuitively consistent with the scope of the data source, that is, the Nuclear Engineering and Technology Journal articles titles between 2013 and 2023, on a high abstraction level.

7.4. Inspecting the t-SNE visualizations of the large dataset

The top TF-iDF terms are prepared for visualization using t-SNE, resulting in 2D plots that are easier to inspect en masse. Firstly, to give a rough idea of the 2D visualization, we plot the top 20 terms ranked by TF-iDF as described in the previous section, to obtain Fig. 11.

The visualization gives more meaning to the list of top 20 terms by representing some aspect of the relationships between them. For instance, we can see that "power" and "plant" on the far left are extremely close, almost overlapping, with some relation to "nuclear", which resides at around a similar distance to "reactor" at the bottom right of the figure. This is an illustration of how the vectors capture semantic relatedness between terms that make up well-known concepts in nuclear domain, even though these terms have been indexed separately. Similar interpretation can be made regarding the proximity of "fuel" and "reactor" at the bottom right. It is also interesting to see that more generic concepts like "simulation", "development", "analysis", or "model", that do not seem to pertain specifically to a particular sub domain of nuclear, are placed together around the centre with relatively similar relations to the other terms, merely based on its central location of the plot. Since absolute coordinates of the points are not of particular importance, rather their relative coordinates to one another, the values on the axis are thus not shown.

In a similar fashion, we can plot a denser version of the same overview of the corpus using the top 100 terms to maintain legibility, to obtain Fig. 12.

Naturally, much higher semantic granularity can be observed with the plot of 100 topics compared to that of just the top 20. On the left, we see the cluster of "safety" "assessment" related topics between "power" "plant" and "accident", near the horizontal axis. Similarly, many well-known multi-term concepts in the nuclear domain have been clearly clustered together. For instance, in the top left quadrant "thermal" and "hydraulic"; "steam" and "generation"; "radioactive" and "waste"; and on the top right quadrant "sodium" and "cooling", "heat" and "transfer"; and in the bottom right quadrant, "spent" and "fuel"; and "monte" and "carlo"; "gamma" and "ray" etc. These numerous obvious cases suggest that the terms, albeit syntactically different, have been vectorized in such a way that even a 2D reductive visualization can capture the semantic relationships between them forming obvious concepts within nuclear knowledge domain.

Following the inspections, an even more detailed landscape of the top 500 terms can be plotted for closer interpretations. However, having

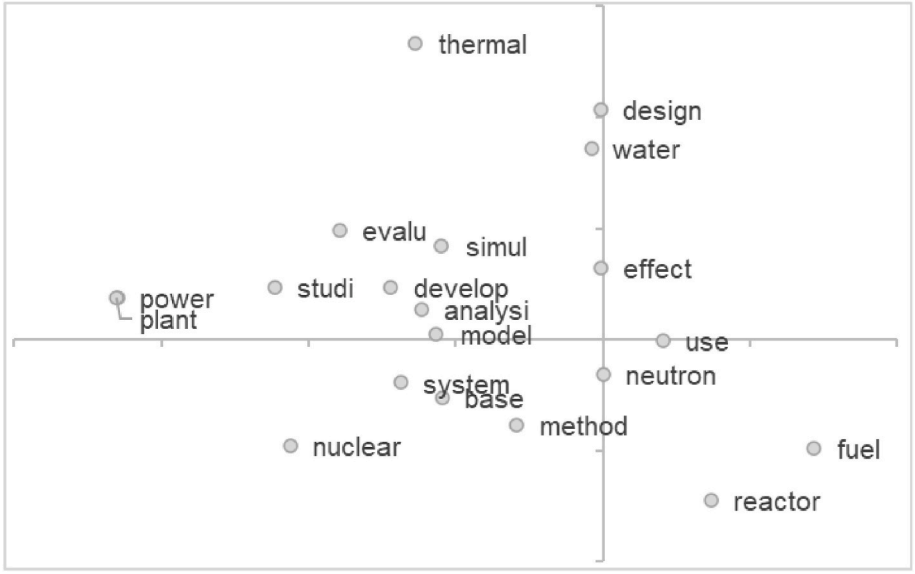


Fig. 11. t-SNE landscape plot of the top 20 topics of the larger dataset.

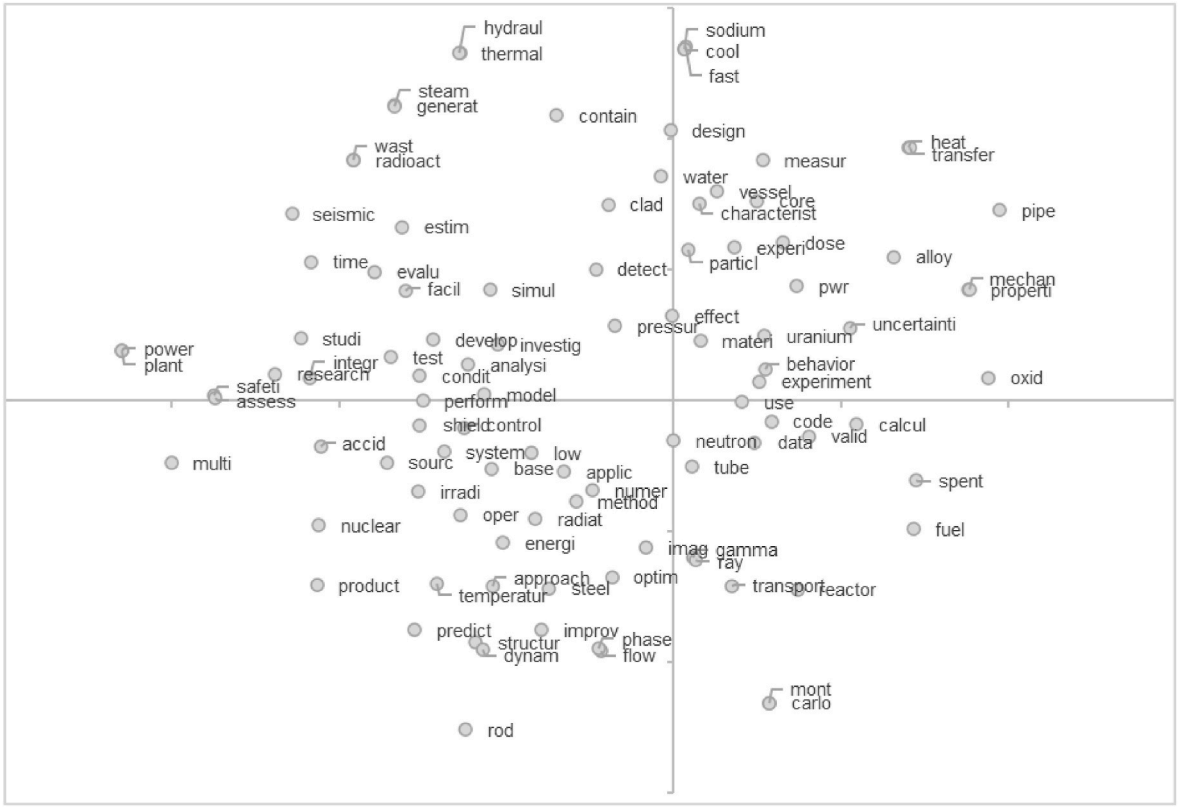


Fig. 12. t-SNE landscape plot of the top 100 topics of the larger dataset.

500 terms is difficult to inspect all at once. Instead, we can zoom in to specific parts to interpret how an even higher granularity of semantic points can be inspected. Magnifying the left half of Fig. 12 along the horizontal axis, gives Fig. 13. The goal is to qualitatively inspect whether intuitively related terms are plotted together by virtue of their vectors, thus the figure enable effective spot checking the landscape to identify recognizable clusters of concepts. The X and Y coordinates and the full 2D visualization of all 500 terms, are available via Mendeley Data (https://data.mendeley.com/datasets/9j6std925r/1) [28].

Towards the left of Fig. 13, we see the cluster of "probabilistic" "safety" "assessment", near its acronym "PSA", which has been tokenized separately, but plotted close by. In the centre of the figure, we see "security" "framework", "risk" and, a bit beneath those, topics such as "accident", and towards its right, "mitigation", "Fukushima" etc. We can interpret that these safety, security, and risk related vectors have been codified so that they are similar enough to be plotted in close proximity to one another. Traversing the landscape in a higher level of detail shows the spectrum of semantic links from one topic to another, compared to high level plots such as Fig. 11 or 12, which represents a more abstract summary of the details. Even though all the vectors exist within the same vector space without any explicit hierarchy, it appears that TF-iDF may also be used as a quantitative method to summarize clusters of other vectors and provide a term that may represent the neighbourhood of many vectors.

Similar qualitative interpretations can be made in other areas of the large landscape of 500 terms. For instance, Fig. 14 shows the magnification of the bottom area of the same 500 term topical landscape.

On the left of Fig. 14, a cluster containing "PET" "resolution", "camera", "imaging", "algorithm" etc. can be observed, and towards its right, topics like "therapy", "gamma", "ray" etc., suggesting the connotation of the medical treatment applications. On the right side of the figure, there are topics about "spent" "fuel", "dry" "storage", and apparently topics about their containers, implied by "cask", and analyses thereof, as per "drop" "impact", and "finite" "element".

Despite the inherent information loss from plotting high-dimensional vectors to a 2D, the obviously related vectors are codified well enough to capture their semantics and thus be plotted within closer proximity. The elucidations of the interpretivist qualitative observations of our discussion should be regarded as just an example of how the inspection can be done. Certainly, those readers who are more familiar with the specialist sub-topics of the nuclear engineering research may see much more connections in the landscape of vectors. Nevertheless, our discussion merely seeks to qualitatively inspect the results of LSA to validate that non-trivial connections between nuclear domain-specific terms and topics can indeed be inferred, for different sizes of datasets and level of semantic detail.

7.5. Scalability and dataset sizes

The scalability of LSA in codifying and enabling sematic linking of nuclear information and knowledge attributes largely to the data sizes that it can handle. Although, there is still little agreement today on the expected size of the LSA corpus or what comprises a large or small corpus [29], LSA has been successfully demonstrated on unstructured datasets of a wide range of sizes, spanning orders of magnitude from just a handful of documents to tens of millions of documents [21,30–33]. In addition, LSA has been specifically shown to perform well on small datasets [34], even those of less than 100 documents [35]. The capability to also perform well on small datasets is an important property of LSA in the nuclear knowledge and competence management use-cases, because this enables those with even very limited datasets to reap the benefits of the semantic inference, compared to other approaches in machine learning, which typically require larger datasets to perform well qualitatively. In this regard, LSA could enable the application of small datasets as a corpus, while maintaining the possibility of extending the corpus overtime. Our empirical explorations of the small ($N = 11$) and larger ($N = 2643$) datasets demonstrate the properties of the semantic inference in both cases, each producing results that indicate the

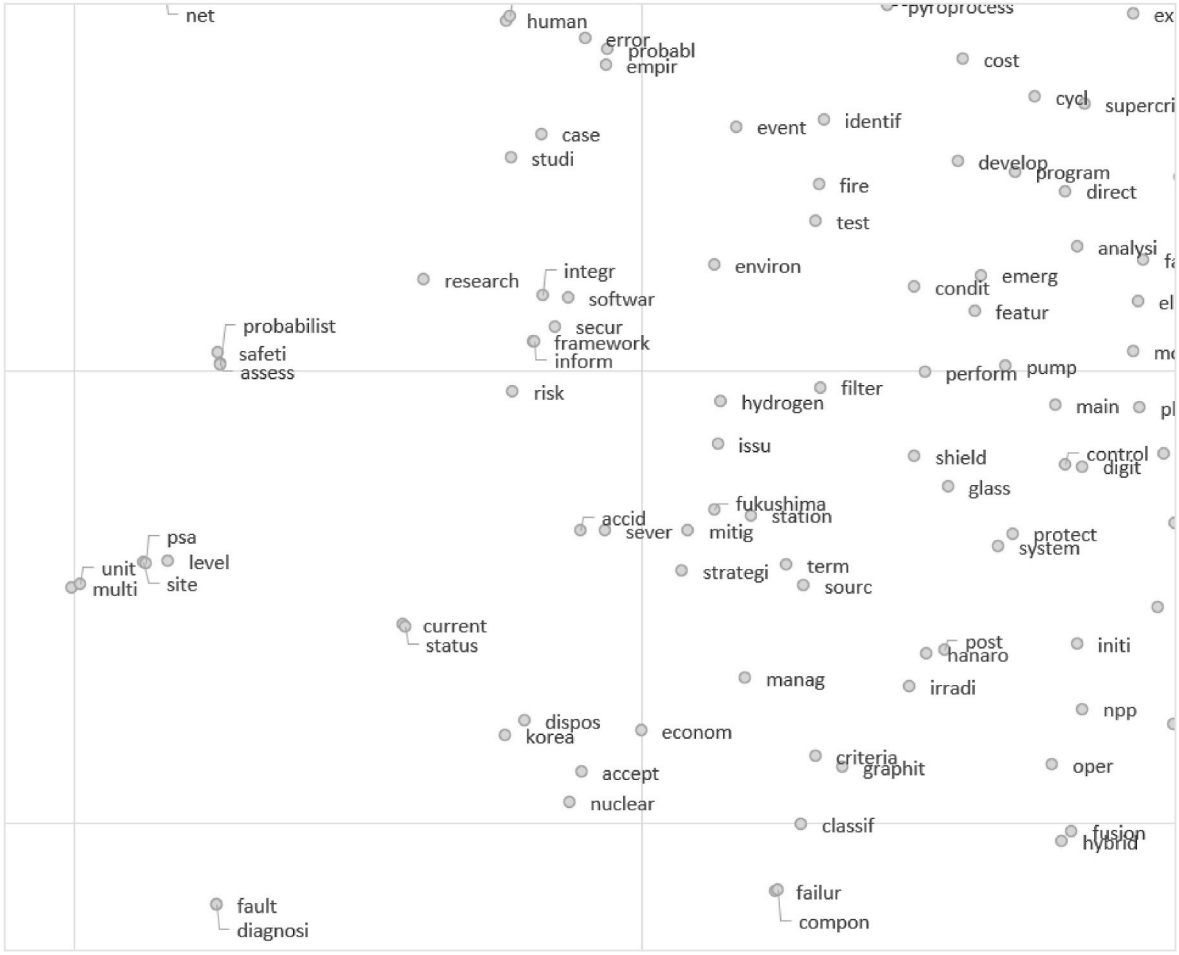


Fig. 13. Left magnification of the t-SNE landscape plot of the top 500 topics of the larger dataset.

semantic vectorization that captures domain specific topical links. These links thus form the semantic structure, which can be utilized in various semantic matching, search or labelling applications for bridging different knowledge communities in an automated and scalable way.

Even though LSA has been demonstrated successfully utilizing a wide range of datasets, the addition or updating of new documents to a corpus can have an implication on the computational requirements of the algorithm in practice. Essentially, there are two ways that addition or updating the corpus can be done, each with its own considerations.

The first way is through the process of "folding-in". This can be described by the process of parsing each new document and weighing them with TF-iDF as if handling a query vector (as per Section 5.5). Thus, new explicit document vectors are transformed to the existing latent semantic space. This enables the new documents' semantic vectors to have the same dimensionality as the existing semantic term and document vectors, thus compatible to comparisons using cosine similarity. This technique is simpler, fast, and less computationally demanding to add new documents to the corpus, since a new LSA vector space need not be recomputed. A drawback to folding-in vectors in this way, when adding new searchable documents, is that new terms, that were not known during the SVD step for the original corpus, are ignored. New terms that do not exist in the old/original corpus will have no impact on the TF-iDF weighting, nor on the semantic structure derived from the original corpus. In practice, this means that the new terms, in the additional folded-in documents, cannot be used as query terms. Although the folding-in process does not account for the new semantic content of the new documents, adding a substantial number of documents in this way will still provide good results for queries, as long as the terms and semantic structures they contain are well represented by the original LSA space. Simply put, if the new documents, that are folded in, are similar to the old/original documents, the retrievals will still be effective.

When the terms and semantic structure of a new set of documents need to be included (to enable queries with new terms), a new term-document matrix and SVD must be recomputed, resulting in a new LSA vector space. Recomputing the new latent semantic structure implies creating a new combined corpus, including existing and new documents, and repeating the whole process. Doing this for larger term-document matrices is computationally costly, requires more time, and for enormous document sets, may not be feasible due to computer memory constraints. Furthermore, the processing time for a query-retrieval cycle of excessively large corpus may become unreasonably long, potentially creating a poor user experience. Nevertheless, documents in the order of millions are readily manageable with today's computational resources, which continue to improve in terms of performance capability. At the same time, a valid strategy for scaling an LSA based retrieval or data linking system could be to combine both methods, that is, recomputing the SVD only occasionally, while folding in additional documents that needs to be searchable quickly.

7.6. Practical value and potential extended studies

The algorithmic investigation demonstrates and deepens our understanding of the empirical behaviours of Latent Semantic Analysis (LSA), which we can extend to aid or replace the manual tasks of information codification and linking. Instead of using predefined

Fig. 14. Bottom magnification of the t-SNE landscape plot of the top 500 topics of the larger dataset.

ontologies/taxonomies, LSA can be used to automatically induce keywords or terms from a corpus of undefined size, consisting of available contents or other textual material in an organization. This reduces the reliance on manpower and thus also addresses the natural inconsistencies of manual work arising from human bias. For instance, manually labelled competence descriptions may result in high variance of the number of keywords used. There is also the problem with synonyms, where human labellers may use different keywords to describe similar content. This has the cascading effect that semantically related keywords need to be linked manually based on domain knowledge. On the other hand, the terms induced and indexed by LSA are vectors in the semantic space, which means that it can be linked with other related vectorized topics via cosine similarity quantitatively. A human moderator is thus not needed to carry out the laborious and time-consuming task of linking related competences or think of synonym keywords during labelling. These have operational value for nuclear knowledge management, for instance, in human resource management and expert competence profiling, modelling, linking, and grouping.

Beside operational viewpoints, there could also be strategic value arising from the technology as suggested by the preliminary study. LSA term vectors can be combined to form document vectors, which act as summaries that describe groups of competences and knowledge communities. These can constitute insights into the larger competence landscape or typology of organizations. Such insights of the knowledge communities of an organization aids decision-making regarding strategic competences. Furthermore, cross-organizational sharing of competence information can open even broader use-cases and implications, that deserve separate in-depth investigations. For instance, LSA and related techniques can be extended to compare the level of alignment (or misalignment) between knowledge/competences of different sectors, for instance, those desired in industry, vs those of educational learning outcomes, vs those promoted by the associations, authorities or the third sector. Dedicated studies can also be done on the effects of different transformations or visualizations of LSA outputs on strategic decision-making amid nuclear ecosystems; or examining the procurement planning processes focussing on assessment of expert competences of bidders on the numerous levels of the supply chain.

Despite the practical relevance of this study, it is important to acknowledge that *technology* is only one of three pillars of an ideal knowledge management, along with *people* and *processes* [4]. Nevertheless, on a principal level, this investigation demonstrates how *technology* can contribute more towards also the soft aspects of knowledge management and enable multidisciplinary collaboration through automated semantic induction from texts. This is a small drop amid the wave of new technological advances in artificial intelligence and data mining, which will continue to reduce the requirement of human effort to collect, organize and manage competence information constantly overtime.

7.7. Limitations

Our methodology and the experimental design do not validate the LSA outputs numerically via other benchmarks or repeated studies. In our case, whether valuable insights can be inferred that is comparable to tacit knowledge, is based on inspection and intuitive interpretation of the LSA outputs from the selected datasets on modelling nuclear competences.

In traditional data/computer science investigations, one would expect that numerical validation is done, with reference to some broad benchmark dataset from other quantitative studies, thus supporting the generalizability of both the algorithm and its outputs. However, given that nuclear knowledge and competence management are still in nascent phases of utilizing advanced technologies, no generally accepted algorithmic benchmarks nor good enough validation dataset exist for our nuclear empirical context. As such, the semantic structures driving the query and retrievals that result from our analyses should not be considered “ground truth”. While this investigation demonstrates the validity and generalizability of the LSA-based method with a small (N = 11) and a larger (N = 2643) dataset, it should be borne in mind that its outputs do not pose as a proof whether it may or may not represent the

same semantic structures of say a particular organization, industry, or sector.

The larger dataset used in this investigation may well represent the scope of knowledge and competences of the nuclear domain, thus creating outputs that are practically useful, intuitively valid, and reasonably transferable. However, it should be borne in mind that, our research experimental design, being rooted in interpretivism, does not seek similar validity, reliability, or generalizability as quantitative research approaches of a positivist worldview.

Due to the tacit, subjective and fragmented nature of knowledge in the nuclear domain, we find it difficult to justify a positivist, quantitative validation philosophy, in which there would still be a burden of proof against bias. Therefore, it has been out of our scope to formulate a separate experimental design for quantitative validation, for instance, with crowdsourced expert respondents, user testing, and to numerically benchmark the performance of LSA in distilling insights. Our outputs only validate that self-evident conceptual and logical associations between nuclear domain-specific concepts can be induced by LSA, that would traditionally have required human work.

8. Conclusion

Our motivation for this study is the challenges of algorithmically modelling competences of nuclear personnel, to enable the effective identification of links between knowledge communities, which in turn, can facilitate steps towards stimulating more multidisciplinary collaboration and transfer of skills. The nuclear field is extremely heterogenous with practitioners from many disciplines, resulting in diverse working circumstances. The number of disciplines involved in the long lifecycles of nuclear facilities and the drawn-out decommissioning and disposal phases are compounded by the heavy regulatory frameworks that govern all operations. This creates an interesting problem to address in terms of nuclear knowledge and competence management, as well as closing the theory-practice gap in knowledge management literature.

The challenge of semantically modelling competences of personnel is largely attributed to the fuzzy, tacit nature of skills and competences, which would have traditionally required human expert cognition to interpret, codify, and manage. Moreover, the communities of practice, as a result of human social behaviour, are self-organizing and constantly changing over time. The constant changes of knowledge communities do not only result from personnel mobility within and across organizations, but also the natural flux of competence divergence of individuals in their careers. Therefore, traditional manual methods for codification of competences (such as tagging of keyword, compilation of taxonomies/ontologies, or identification of links between related topics) are not feasible with large datasets and require regular updates to the codification over time.

This investigation demonstrates in detail the capabilities and limitations of the LSA technique in automatically identifying and linking competences semantically, induced from typical competence descriptions in abundant natural language texts. The observations and interpretations of the results of this investigation improve our understanding of how outputs from domain-specific textual datasets, of varying sizes describing personnel competences, can be codified automatically and retrieved, while capturing aspects of semantics that is implicit to human interpretation within the respective domain.

In essence, the theoretical basis that brings together the semantics (tacit) and the computations (explicit), is that semantically related words/terms usually tend to appear together in a body of text and co-occur by some consistent degree of separation. This complex network of probabilistic relationship structures is identified through singular value decomposition. Following that, dimensionality reduction removes patterns that are considered "noise", which introduces eccentricity to the desired semantic structures. This property of LSA is fundamental in effectively addressing the fuzziness of natural language descriptions of competences attributed to individual nuclear personnel, from which inferences can be made about the communities of practice that they constitute.

Even with the small dataset of this investigation (N = 11), the latent semantic patterns can exhibit traits of artificial intelligence, inducing synonyms and semantic relationships. This changes the typical assumption of the need for "big data" to yield such results, and in our case a variety of queries can already be made with sensible, explainable retrievals. A small dataset that can be easily interpreted, demonstrates the in-depth mechanisms of LSA. On the other hand, the larger dataset (N = 2653) in our analysis poses a realistic scale, scope, and depth of semantic content akin to that of descriptions of nuclear domain specific knowledge and competences. The visualizations of LSA outputs for the larger dataset using t-SNE also show intuitive links upon interpretation on varying levels of abstraction. Given the capability of LSA to handle a wide range of dataset sizes effectively, it is feasible to have even larger input training datasets and thus larger vocabulary of terms. The variety of queries could be augmented with growth in the training data.

We validate the potential of LSA as an underlying technique for different applications related to the management of knowledge communities that enables operational and strategic decision-making, linking and identifying related nuclear competences across silos. Above all, the technique is automatic and utilizes existing abundant textual data, without the need for manual input and regular updating of competence data, as typically required in traditional knowledge management software portals.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] O. Matselyukh, C. Chenel-ramos, M. Ceclan, Retaining Critical Competences in Nuclear Energy Sector: National Initiatives and Best Practices, Instruments and Tools, 2015.
[2] L. Edvinsson, M. Malone, Intellectual Capital, Harper Business, New York, New York, USA, 1997.
[3] E. Wenger, R. McDermott, W. Snyder, Cultivating Communities of Practice: A Guide to Managing Knowledge, Harvard Business School Press, 2002.
[4] L.A. Petrides, T.R. Nodine, Knowledge Management in Education: Defining the Landscape, 2003. Half Moon Bay.
[5] C. Bratianu, I. Orzea, Organizational knowledge creation, Manag. Mark. Challenges Knowl. Soc. 5 (2010) 41–62.
[6] G. Von Krogh, The communal resource and information systems, J. Strat. Inf. Syst. 11 (2002) 85–107.
[7] S.L. Pan, D. Leidner, Bridging communities of practice with information technology in pursuit of global knowledge sharing, J. Strat. Inf. Syst. 12 (2003) 71–88.
[8] G. Bell, F. Lai, D. Li, Firm orientation, community of practice, and Internet-enabled interfirm communication: evidence from Chinese firms, J. Strat. Inf. Syst. 21 (2012) 201–215.
[9] J. Kietzmann, K. Plangger, B. Eaton, K. Heilgenberg, L. Pitt, P. Berthon, Mobility at work, J. Strat. Inf. Syst. 22 (2013) 282–297.
[10] IAEA (International Atomic Energy Agency), Knowledge Management and its Implementation in Nuclear Organizations, IAEA Nucl. Energy Ser. No. NG-T-6.10, 2016.
[11] IAEA (International Atomic Energy Agency), Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management, IAEA Nucl. Energy Ser. No. NG-T-6.15, 2021.
[12] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.
[13] S.T. Dumais, Latent semantic analysis, Annu. Rev. Inf. Sci. Technol. 38 (2004) 188–230.
[14] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, Waltham, USA, 2006.
[15] S.T. Dumais, O.M. Way, LSA and information retrieval: getting back to basics, in: T. K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (Eds.), Handb. Latent Semant. Anal, Lawrence Erlbaum Associates, Mahwah, NJ, 2007.
[16] C. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, 2008.
[17] M.A.F. Ragab, A. Arisha, Knowledge management and measurement: a critical review, J. Knowl. Manag. 17 (2013) 873–901.
[18] K.M. Eisenhardt, M.E. Graebner, Theory building from cases: opportunities and challenges, Acad. Manag. J. 50 (2007) 25–32.

[19] W.J. Creswell, D.J. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approach, SAGE Publications Ltd, 2018.
[20] European Commission, Petrus III. https://cordis.europa.eu/project/rcn/109649_en.html, 2015. (Accessed 25 April 2018).
[21] V. Kuo, Latent Semantic Analysis for Knowledge Management in Construction, Aalto University, 2019.
[22] T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch, Handbook of Latent Semantic Analysis, 2007.
[23] S.S. Kulkarni, U.M. Apte, N.E. Evangelopoulos, The use of latent semantic analysis in operations management research, Decis. Sci. J. 45 (2014) 971–994.
[24] A. Sidorova, N. Evangelopoulos, J.S. Valacich, T. Ramakrishnan, Uncovering the intellectual core of the information systems discipline, MIS Q. 32 (2008) 467. A20.
[25] N. Evangelopoulos, X. Zhang, V.R. Prybutok, Latent semantic analysis: five methodological recommendations, Eur. J. Inf. Syst. 21 (2012) 70–86.
[26] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[27] M.F. Porter, An algorithm for suffix stripping, Progr. Electron. Libr. Inf. Syst. 14 (1980) 130–137.
[28] V. Kuo, Nuclear Engineering and Technology Journal Titles Dataset and LSA Outputs, Mendeley Data, vol. 1, 2023. https://data.mendeley.com/datasets/9j6std925r/1.
[29] S.A. Crossley, M. Dascalu, D.S. McNamara, How important is size? An investigation of corpus size and meaning in both latent semantic analysis and latent dirichlet allocation, in: Proc. Thirtieth Int. Florida Artif. Intell. Res. Soc. Conf., 2017, pp. 293–296.
[30] S. Sarkar, A. Dong, J. Gero, Learning symbolic formulations in design: syntax, semantics, and knowledge reification, Artif. Intell. Eng. Des. Anal. Manuf. 24 (2010) 63–85.
[31] N. Evangelopoulos, Latent semantic analysis, Wiley Interdiscip. Rev. Cogn. Sci. 4 (2013) 683–692.
[32] A.M. Olney, Large-scale latent semantic analysis, Behav. Res. Methods 43 (2011) 414–423.
[33] R.B. Bradford, An empirical study of required dimensionality for large-scale latent semantic indexing applications, in: CIKM '08 Proc. 17th ACM Conf. Inf. Knowl. Manag., 2008, pp. 153–162.
[34] Y. Hong, H. Xie, G. Bhumbra, I. Brilakis, Comparing natural language processing methods to cluster construction schedules, J. Construct. Eng. Manag. 147 (2021).
[35] T. Cvitanic, B. Lee, H.I. Song, K. Fu, D. Rosen, LDA v. LSA: a comparison of two computational text analysis tools for the functional categorization of patents, in: Int. Conf. Case-Based Reason., 2016.