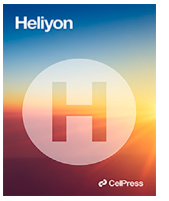
Heliyon 10 (2024) e29004

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Exploring hidden pathways to sustainable manufacturing for cyber-physical production systems

Gianfranco Pedone *, József Váncza, Ádám Szaller

Research Laboratory on Engineering & Management Intelligence, Centre of Excellence in Production Informatics and Control, Institute for Computer Science and Control, Kende u. 13-17, Budapest, 1111, Hungary

ARTICLE INFO

Dataset link:
github.com/SZTAKI-EMI-PG-public/RS4SCPPS

Keywords:
Sustainability
Cyber-physical production system
Text analysis
Semantic similarity

ABSTRACT

Future manufacturing scenarios will likely be built around cyber-physical production systems. To succeed, this new manufacturing paradigm will also have to comply with the golden rule of sustainability. However, the concept of sustainability as defined in a number of high-level policy documents and recommendations requires disambiguation. The paper introduces HITECS, a novel, context-independent text analytics methodology for hidden correlation analysis in documents. HITECS is based on the assumption that there is a strong link between a concept and the words implicitly chosen to explain it. The analysis is based on the combination of bare words frequency and cosine similarity, excluding trivial, first-level terms (titles, keywords, and definitions). Processing a corpus of generally accepted documents related to various definitions and requirements of sustainability unfolded their hidden correlations and some common key concepts. These results indicate that terms like *access, inclusion, global, change*, together with others like *resource, share*, and *integration*, are among leading concepts in the high-level documents discussing the requirements of sustainability. A similar analysis in the domain of cyber-physical production systems shows strong conceptual overlaps but also gaps indicating pathways for future research and actions.

1. Introduction

The impact of the fourth industrial revolution on the social and natural environment is considered significant and far-reaching, even though the interactions of the human, natural and manufactured assets are less understood, extremely complex and partly unpredictable. The challenge of *sustainability* is how to use all these capital in the most efficient way, also complying with their absolute limits, in the service of the well-being of society. Due to its direct relation to manufactured assets, production engineering has a special role and responsibility in this respect. Our main research interest is whether and how present days' so-called *cyber-physical production systems* (CPPSs) operating in the fabrics of society and the natural environment can respond to this challenge [1,2].

There is a common understanding that contemporary information and communication technologies (ICT) employed in CPPSs can be technological enablers of sustainable production. However, no informatics solution can help answering the key question, namely, how to handle the intrinsic asymmetry between short-term rewards of the markets and long-term stewardship of the natural and social resources. How can we align requirements of industrial efficiency and competitiveness with those of sustainability? The resolution of this dilemma certainly calls for new approaches, mindsets, and institutions. We are convinced that beyond — and before

* Corresponding author.
E-mail addresses: gianfranco.pedone@sztaki.hu (G. Pedone), jozsef.vancza@sztaki.hu (J. Váncza), adam.szaller@sztaki.hu (Á. Szaller).

https://doi.org/10.1016/j.heliyon.2024.e29004
Received 11 January 2024; Received in revised form 26 March 2024; Accepted 28 March 2024

Available online 3 April 2024

— providing particular answers of technology, production engineering has to contribute to the formation of economic, social and legal requirements and institutions that define pathways to a more cooperative world of production and correct the route what might go wrong with the first industrial revolution [3].

The ultimate goal of this work is to identify pathways of CPPS towards sustainable manufacturing. However, efficiency and competitiveness are relatively well-known and well-defined terms for production, in contrast to the notion of *sustainability* [4]. Indeed, sustainability is a problem we discuss in many contexts without having a clear, common understanding and definition of its notion. As it is poignantly discussed by [5], the however well-intentioned conceptual variants of sustainability can only distract us from defining the right goals and from finding the right pathways to attain them. It is a "wicked problem" due to the divergent values of shareholders. Hence, its definition, circumscription and solution without openly acknowledging conflicting views [6] and relying on practical wisdom [7] seems to be hopeless.

We see at least three factors blurring a clear understanding of sustainability, even within the narrower context of production. First, there is still a huge gap between the world of research and the one of policy-makers and people education [8]. Certainly, a discrepancy is in the way how common problems are discussed and interpreted. Secondly, even by adopting the same terminology, there seems to be a basic misunderstanding of the fundamentals of the addressed issues. This is true also for the related concept of *circular economy* [9,10] and its implications [11]. This is addressable to the fact that these high-end, main-streaming concepts are explained and approached in very different manners, taking into account very different priorities and economic, societal and environmental conditions. Finally, what various stakeholders say is often not essentially what they really mean on the basis of their overall proposed argumentation.

Hence, the paper centres around the following questions: Is it possible to find, if any, inner concepts at the heart of sustainability in such a variety of divergent domain-specific discussions? And more importantly, is it possible to identify specific and common directions for actions? Are these concepts relevant also in the domain of CPPSs? Can we identify conceptual overlaps and gaps between sustainability and cyber-physical production, and if so, can these findings help set the basis to the advancement of CPPSs towards sustainable manufacturing?

The answers are pursued by *text analytics* (TA) which can help disambiguate the distance of argued concepts. TA is one of the most critical ways of analyzing and processing unstructured data, typically texts which represent nearly 80% of the world's data. Today a majority of organizations and institutions gather and store massive amounts of textual data in dedicated data warehouses and cloud platforms. This data continues to grow exponentially and is collected from more and more sources. The result is a situation in which it becomes a challenge for companies and organizations to store, process, and analyze vast amounts of textual data with traditional tools. According to [12], TA (also referred to as text mining or text data mining) is the process of deriving actionable information from text, which incorporates a multidisciplinary approach, spanning from information retrieval to machine learning, statistics, and computational linguistics. TA deals with natural language texts either stored in semi-structured or unstructured formats, being part of a much wider family of natural language processing (NLP) methods [13].

In this research paper we present *HITECS* (HIdden TExtual Correlation Seeker), a novel methodology and tool whose aim is to discover *hidden correlation concepts* among a corpus of documents. In this case, HITECS is addressing sustainability and cyber-physical production. Specifically, we depart from high-level sustainability related documents from world-wide, global organizations and focus on identifying the main implications, both generally and on CPPSs. Selected topics focus on sustainable development, circular economy, information and communication technologies, and cyber-physical production systems, but the methodology can be applied to any document and domain containing the concepts of interest.

The expression *hidden* refers to the fact that a similarity analysis will be executed by excluding all terms contained in primary queries and definitions: article keywords, document titles, scopes and domains of the examined documents in the corpus. This is the fundamental aspect of HITECS when compared to traditional text analytics techniques: the latter usually produce *obvious* correlation high-scores due to the fact that the same keywords either characterize the queries underlying the identification of a concept *and* are eventually contained also in the match-making logic of the correlation analysis. Our objective was, on the contrary, to examine the bare (sanitized and normalized) version of the documents and measure the level of conceptual distance emerging from each other under these conditions. In other words, HITECS will try to discover which concepts are implicitly (or unconsciously) leveraged by scientists, policy makers, regulators, and opinion leaders when discussing sustainability from their own perspectives.

With the application of this novel methodology, we will try to find answers to questions like: What is the real meaning of argumentation in actions and plans for a sustainable manufacturing? What do regulators and experts really have in mind when discussing about sustainability? How do they really intend sustainability in terms of actionable steps? Are there any common pillars for sustainability in the different application domains? What do regulators and members of the scientific society actually refer to when trying to define new approaches for a sustainable future?

In the sequel Section 2 shortly summarizes related works. Next, Section 3 presents the principles and workflow of the HITECS methodology. This section gives also an overview of the document corpus selected for analysis, and of the most necessary technical details of processing them. Section 4 is dedicated to the explanation and interpretation of the results, highlighting the main implications of our finding also to CPPS. Validity and operative limitations of HITECS are summarized in Section 5, and conclusions are drawn in Section 6. An Appendix containing further computational details on HITECS is also available in the paper. All master data, parameters and sources utilized in HITECS analysis are made publicly accessible via a GitHub repository.[1]

[1] https://github.com/SZTAKI-EMI-PG-public/RS4SCPPS.

2. Related works

Despite a visionary misalignment on sustainability, countries and world organizations are moving towards a more *sustainable* development. Examples of definition, criteria, indicators and empirical analysis can be found in [14]. Institutional pressure for sustainable supply chain management and circular economy is also being applied, as reported for example in [15].

Most research studies typically summarize the state of the art by systematic bibliographical search using standard terms, followed by document filtering and analysis. Such an influential study presents, for example, various approaches to business model innovation in the services of circular economy and sustainability [16], or a more recent one discusses how circular economy works in various industrial settings [17]. Alternatively, expert opinion can be collected also by directed surveys. Production engineering early realized its responsibility in sustainable manufacturing and investigated the potential impact of the life cycle engineering (LCE) approach on attaining the United Nations' Sustainable Development Goals (SDGs). Specifically, via interactive sessions experts were consulted on whether and how LCE can contribute meeting these goals. Altogether 33 UN SDG targets were considered relevant to the manufacturing industry, and the study identified the top 10 SDGs that life cycle engineering could contribute to undertake [18]. While admitting the importance of such approaches which result in condensed surveys and policy papers, here we suggest an alternative, orthogonal way to elaborating key concepts and their relations in the sustainability-oriented scientific and social discourse.

There is also the crucial aspect of sustainable profitability, which needs to be taken into account in the definitions and organizational practice of sustainability, as evidenced in studies already from the past decade [19]. There are research studies that try to define and identify *strongly sustainable* product-service systems, concentrating on aspects such as access, substitution, systemic dematerialisation, territorial anchoring and sufficiency [20]. Current reality, future potential and challenges for innovative products and services in the light of a sustainable societal development can be found in [21]. A very interesting literature review presented in [22] focuses on how Industry 4.0 is defined from a socio-technical perspective and how much sustainability it includes. As for CPPS definition and approaches, [1] provide a clear and comprehensive analysis of cyber-physical systems in manufacturing. Anyway, sustainability might not be only a matter of materials, environment, resources and optimization. Recent publications have tried to assess the existence of a link between sustainability and spirituality, as the basis for a more ethically correct world [23].

HITECS is *not* another method for literature review, but a TA methodology for programmatically discovering hidden concepts gluing together a corpus of documents, whose correlation is analysed by eliminating all the terms explicitly categorising them and that have been intentionally assigned by the authors to the document (title, keywords, definitions, meta-data, etc.). Typical structured literature review (SLR) like in [24] uses a standard systematic approach by means of manual searches and human-driven source-processing at all stages. Moreover, SLR methods usually leverage the *same set* of query terms both in the corpus generation and in the definition of an ad-hoc categorization framework necessary for the analysis of the specific method's outcome [25,26].

In HITECS only the pre-processing phase requires some basic, initial, manual configuration when terms utilized for the generation of the corpus (Step 1 on Fig. 1) are excluded from the next stages of the methodology and from correlation analysis. In other words, HITECS is context-independent.

3. HITECS methodology

The main novelty of HITECS approach relies on a different combination strategy of term frequency, document vectorization (exclusion of all trivial, first-level key-words) and cosine similarity. So-called *term frequency* and the related *inverse document frequency* (TF/IDF) are well-known techniques used to quantify words in a set of documents. Generally, a score is computed for each word to signify its importance in the document and the overall corpus. This is a widely used technique in information retrieval and text analytics, by means of which it is easier for an "artificial system" to understand textual data in the form of numerical value. For this reason, we need to *vectorize* all the text from all the selected documents, so that it is interpretable by programs. This is very likely the same technique used by search engines, for example, which maintain a fixed representation of all the documents available on the web and find the relevance of the user queries with all of the documents, rank them in the order of relevance and show the top k documents. All this process is done using the vectorized form of queries and documents.

3.1. Principles of the methodology

The key principles of our analytic methodology are the following:

i) *First-level terms exclusion*. The so-called facade terms (queries, keywords, titles, domains and scopes) related to the investigation are not considered in the set of correlating words, as they would produce trivial matching scores in the analysis of contents. The vocabulary of the corpus is built as the union of all terms identified starting from the second level in the corpus.

ii) *Term relevance not uniqueness*. The aim is not to discover how unique a term is in the corpus of documents, but how relevant the term is for a specific document, once eliminated all the irrelevant stop-words (conjugations, conjunctions and unwanted terms from previous point). In the light of these premises, we calculate the term frequency (TF) for a document, whereas the document frequency (DF), which is the number of documents containing the word, will be calculated only in case of correlation. This information is implicitly leveraged by the *cosine similarity* between two term-vectors.

iii) *Document normalization*. TF measures the frequency of a word in a document, but this highly depends on the length of the document and the generality of the word itself. Moreover, we cannot say that a longer document is more important than a

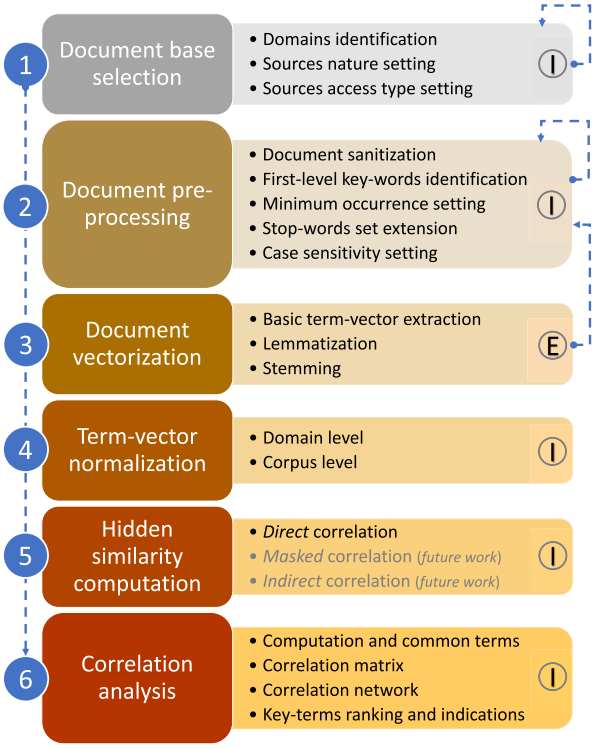


Fig. 1. Hidden correlation discovery methodology.

shorter one. For this reason, we also perform normalization (through linear scaling) of document term-vectors; taking into account the ratio between a document's weight and the "heaviest" document available in the corpus.

iv) *Term-vector parameterization*. We define a practical (dynamic) limit in the number of eligible (i.e. not contained in the exclusion - also called stop-word or blacklist) words to be considered as relevant (used at least two times in the document). In our case, the limit is one hundred. It is plausible to think that the first one-hundred, not "accidental" (i.e. $frequency > 1$) words in the bare document (i.e. sanitized from irrelevant sections) can provide an acceptably meaningful footprint of its term-vector. This is a computational simplification, nevertheless, our approach works properly also with term-vectors of arbitrary length (i.e. not need to be of the same size);

v) *Cosine similarity enhancement*. We combine word-matching with cosine similarity (CS), as the former might not be reliable in the case of long queries on the documents [27]. Cosine marks all the documents as vectors of TF tokens (weights) and measures the similarity in cosine space (the angle between the vectors). CS ranges between zero (0) and one (1), inclusive and will actually indicate the level of correlation (conceptual distance) between documents (section 3.6).

All in all, the above principles differentiate our analytics method from traditional TF/IDF techniques and define the basis for a novel approach for detecting hidden concepts and relations in texts which deal with "wicked problems", such as sustainability.

The methodology is organized into six major phases, as depicted in Fig. 1. Step numbers indicate the sequence order; dotted, self-referencing arrows refer to iterative steps, whilst the labels "*I*" and "*E*" indicate whether the phase was developed internally or provided through an external service. Each step is briefly explained in the following. For further details, please, refer to the Appendix A.

3.2. Document base selection (step 1)

The methodology introduced in this paper is based on a highly diversified bibliography, which better supports final conclusions on a footprint of a much wider community. Sources were selected not only from primary publication main-streams, such as scientific journals and economic forums, but also from publicly available (online) document repositories, blogs and releases of opinion-leading organizations and regulators.

The dimension of the investigated corpus aims at the purposes of HITECS and guarantees an acceptable compromise between the minimum coverage of the domains and the manual effort necessary in the pre-processing phase of the methodology. In compliance with the prerequisites of HITECS, the corpus is still sufficient to produce evidence of unbiased, hidden correlations. Selected items in each domain derive from SE prioritization strategy in their ranking algorithm, which is preferably based on independent citations and references.

Table 1
Corpus of the analysed documents.

Domain	Scope	Doc ID	References
Sustainability base	Overall definition, history, evolution	SB_gen	[28–30]
	UNESCO	SB_une	[31]
	NATO	SB_nat	[32]
	EU	SB_eu	[33]
	UN	SB_un	[34]
	UNICEF	SB_uni	[35]
Sustainability pillar	Environment	SP_env	[36,37]
	Society	SP_soc	[38]
	Economy	SP_eco	[39]
	Culture	SP_cul	[40]
Cyber-physical Production System	Overall, design, implementation, classification, comparison	CPPS_gen	[41–43]
Information and Communication Technology	Digitalization	ICT_dig	[44]
	Artificial Intelligence	ICT_ai	[45]
	Blockchain	ICT_blc	[46]
	Cloud	ICT_clo	[47,48]
	Big Data	ICT_big	[49]
	Hardware	ICT_har	[50,51]
Circular Economy	General	CE_gen	[52]
	114 Definitions	CE_def	[9]
	Partnership building	CE_par	[53]
UN Agenda 2030 - 5P	General	UN5P_gen	[54]
	People	UN5P_peo	[55]
	Planet	UN5P_pla	[55]
	Prosperity	UN5P_pro	[55]
	Peace	UN5P_pea	[55]
	Partnership	UN5P_par	[55]
UN Sustainable Development Goal	No poverty	UNSDG_01	[56]/goal1
	Zero hunger	UNSDG_02	[56]/goal2
	Good health and well-being	UNSDG_03	[56]/goal3
	Quality education	UNSDG_04	[56]/goal4
	Gender equality	UNSDG_05	[56]/goal5
	Clean water and sanitation	UNSDG_06	[56]/goal6
	Affordable and clean energy	UNSDG_07	[56]/goal7
	Direct work and economic growth	UNSDG_08	[56]/goal8
	Industry, innovation and infrastructure	UNSDG_09	[56]/goal9
	Reduce inequalities	UNSDG_10	[56]/goal10
	Sustainable cities and communities	UNSDG_11	[56]/goal11
	Responsible consumption and production	UNSDG_12	[56]/goal12
	Climate action	UNSDG_13	[56]/goal13
	Life below water	UNSDG_14	[56]/goal14
	Life on land	UNSDG_15	[56]/goal15
	Peace, justice and strong institutions	UNSDG_16	[56]/goal16
	Partnerships for goals	UNSDG_17	[56]/goal17

Selected domains focused on sustainability (environment, society, economy and culture), CPPS, circular economy, ICT (artificial intelligence, big data, cloud, block-chain and hardware) and the 17 United Nations SDGs. As for the type of sources access, we have targeted open web and open-access journals, as well as publicly accessible repositories. The list of all selected documents, their application domain, their scope, the assigned IDs, as well as their references are reported in Table 1. References related to the UN SDGs [56] (last row in the table) are reported with the simplified schema *SDGURLBase*[2]/goal**X** (with X ranging from 1 to 17), as they all share the same URL (Unified Resource Locator) root.

3.3. Document pre-processing (step 2)

Document sanitization is a crucial aspect of the methodology and was necessary to ensure that only the *relevant* parts of a document are analyzed during the computations, removing all the (meta)data inherent in the document structure or the secondary text. Document sanitization was a semi-automated process, considering the various nature of documents accessed.

First-level keywords exclusion was the next fundamental part of this phase. These terms, characterizing the logic of document querying and identification in the corpus (previous point), need to be excluded from the correlation analysis. This is the very sense

[2] SDGURLBase = https://sdgs.un.org/goals.

of this research work: eliminating trivial correlations deriving from obvious textual match-making (for example: "*energy, direct, work, economic, growth, industry, innovation, infrastructure*", and many more). The union of all first-level keywords with the stop-word list of the English language provided the final set of terms to be ignored in the vectorization of a document (next phase). The full list of these terms can be accessed on our GitHub project folders.[3] The minimum number of occurrences for a term to be eligible in vectorization was set to 2, in order to avoid the term's *casualty* in the text. The application of lowercase settings closes this step.

3.4. Document vectorization (step 3)

The basic form of document term-vectors was obtained via an external SaaS.[4] Each term-vector can be represented as a *row* in the matrix below (1):

$$W_{n\times m} = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & \cdots & t_m \end{matrix} \\ \begin{pmatrix} w_{1,1} & w_{1,2} & w_{1,3} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & w_{2,2} & \cdots & w_{2,m} \\ w_{3,1} & w_{3,2} & w_{3,3} & \cdots & w_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & w_{n,3} & \cdots & w_{n,m} \end{pmatrix} \end{array} \begin{matrix} doc_1 \\ doc_2 \\ doc_3 \\ \cdots \\ doc_n \end{matrix} \tag{1}$$

where $w_{i,j} \geqslant 0$ is the weight (frequency) of the j-th term in the i-th doc of the corpus, n and m is from (A.5). Where a term t_j is not present in a doc_i then $w_{i,j} = 0$. The term-vector base targets the selection of the 100 most relevant (frequent) words in a document. Term-vectors were eventually lemmatized. Lemmatization aims at reducing the inflectional forms and sometimes derivationally related forms of a word to a *common base* form (or root form). For our investigations we have utilized an external open-source lemmatization service.[5] Once lemmatized terms-vectors underwent the process of stemming, which is similar to lemmatization but aims at reducing the inflectional forms of terms by a crude heuristic process that truncates the ends of words for the removal of derivational affixes. The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is Porter's algorithm [57]. For our investigations we have utilized an external, open-source stemming service,[6] based on Porter's outcomes.

3.5. Document normalization (step 4)

Analysed documents had different lengths, so selected terms' weights initially reflected this. We have set up a normalization process in order to scale the weight of the documents and so to guarantee the comparability of documents and the meaningfulness of the similarity computations. The linear scaling rate was calculated against the heaviest document available in the context of interest (first at the local, and then at corpus level).

It is important to highlight that the global scaling (at the corpus level) has to follow the local ones, in order to preserve the initial consistency and distance of term weights. Local normalization consisted of the application of the linear scaling to multiple documents of the same scope in a domain (like *sustainability base definitions* or *ICT cloud*, for example. Refer to Table 1 for the complete list). This step also produces the union of the scope-specific term vectors into a single, scaled and merged one, whose elements are the first one hundred heaviest terms selected among all the terms after the scaling (all details of the pre-processing are available on our cloud repository).

3.6. Cosine similarity (step 5)

CS is an $n * n$ diagonally symmetric matrix, whose computation is based on formula reported in (A.5) and whose structure is as follows:

$$C_{n\times n} = \begin{pmatrix} 1 & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\ c_{1,2} & 1 & c_{2,3} & \cdots & c_{2,n} \\ c_{1,3} & c_{2,3} & 1 & \cdots & c_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{1,n} & c_{1,n} & c_{3,n} & \cdots & 1 \end{pmatrix} \tag{2}$$

where $c_{i,j} \in [0, 1]$ and n is from (A.5) after documents normalization at global scope. As correlation is a symmetrical property, we will obtain that $c_{i,j} = c_{j,i}$ where $i \neq j$, 1 otherwise (reflexive property on the diagonal).

A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no *conceptual* match. The closer the cosine value is to 1, the smaller the angle is and, therefore, the greater the match is between the two vectors. CS will actually indicate the level of correlation (conceptual distance) between documents.

[3] https://github.com/SZTAKI-EMI-PG-public/RS4SCPPS.
[4] https://tagcrowd.com/.
[5] https://seohorsesense.com/free/lemmatization.php.
[6] https://textanalysisonline.com/nltk-porter-stemmer.

3.7. Correlation analysis (step 6)

In this final phase, the conceptual relation between two documents, as well as the ranking of the most occurring correlation terms were analyzed and a possible interpretation was proposed. CS is a practical tool for describing simple relationships, without making any statement about cause and effect, but only about the evidence that two documents have hidden concepts in common, which may unfold and strengthen the convergence of vision on sustainability pathways. All computations from points 4 to 6 in Fig. 1 are available as JAVA sources on GitHub.[7] *DC* results are reported in Section 4.

All $c_{i,j}$ values from (2) have been visually organized in Fig. 2 (for simplicity, only one side of the matrix has been reported, as it is symmetric). The correlation matrix contains a heat map, too, in which darker cells indicate a higher level of correlation between two documents. Domains borders in the corpus have been highlighted with dotted boxes, for a more immediate comparison of outcomes among inner scopes.

In addition to the correlation matrix, a *correlation network* will provide a graph-based view of the links emerged in direct correlation (DC) computations (Fig. 3). Documents represent the nodes of the peer-to-peer network, whereas edges reflect the actual existence of connections for a specific document pair, emphasizing the highest correlation for a document with a thicker stroke.

Finally, all the correlating terms were also organized according to their ranking and frequency weights in correlations. For each correlated document, we analyzed i) the number of occurrences of common terms; ii) a term's overall weight (frequency) from correlations containing it; as well as iii) the average correlation weight of the term. A 3-dimension ranking list for the first 40 most frequent correlation terms was also generated, which provided, for each dimension, indications on their relevance and *invariance* when traversing the dimensional rankings.

4. Results and discussion

In this section, we highlight and discuss the most relevant outcomes (and their implications) of our methodology applied to the selected corpus.

4.1. Document dimension

Illustration in Fig. 2 is an aggregated representation (heat-map and DC scalar matrix) of the document correlations. Each DC cell is identified by a value and a colour intensity: the higher the DC score is, the darker is the cell colour in the scale range. The DC matrix has been divided into quadrants (dotted lines with domain label) for an easier identification and comparison of results. Each quadrant contains the DC values computed for scopes between two different domains of investigation. Quadrants on the matrix diagonal, on the contrary, reports the DC related to scopes within the same domain.

As a general comment in our investigations, it has trivially resulted that domains in which documents were provided by the same organization body (i.e. UNSDG or UN5P, that is United Nations in the core) evidenced a much higher level of correlation (even in the absence of "front-end" terms). This is probably due to the fact that, regardless of the size and the worldwide level of organizational distribution, there exists some sort of enterprise global convergence or orientation (dependency), which leads, in the long run, to the adoption of a unified portfolio of expressions and challenges vision within an organization.

Let us proceed with order. For each of the domains in the corpus (Table 1) we summarize hereafter the most relevant outcomes in their correlations. Value scores from the DC matrix have also been reorganized into a DC network (Fig. 3), with the intention to offer a more immediate visual understanding of the connections between documents in the corpus. The highest DC value between two documents (nodes on the circle) has been emphasized by means of a thicker grey chord.

4.1.1. Sustainability basics

It evidences a higher DC between documents of the domain in relation to the general understanding of sustainability, even though not from the same regulator (UN, UNESCO, NATO). This implies that there are concepts which underpin and support the common definition of the objectives and their discussion. It is very interesting to notice that the DC with CPPS is relatively low, ranging from an encouraging 0.14 (and 14 words in common) with *CPPS_gen* to a very poor 0.01 (4 terms in common) from *SB_eu*. This can be interpreted as the (critical) lack of common vision and understanding of sustainability between technicians of the CPPS and the definers of sustainability in different domains.

Also, ICT shows a relatively low level of DC with sustainability basics, even though there are specific scopes that produced an important level of correlation, like digitalization in general, cloud technology, big data and the increasing hardware requirements for computations (primarily, GPUs, data warehouses and communication networks). These scopes are crucial for the definition of a mutually understandable notion of sustainability, and it is clear that they play a fundamental role in the measurement of ICT's overall environmental footprint.

The Circular Economy domain, despite its innovative and promising business appeal, shows levels of DC similar to ICT's (higher DCs are scored with *SB_gen*, only). This is a very clear symptom of the fact that once eliminated all the common buzz-words that glue together the surface of scientific and business discussions, very few concepts remain as backbone of the common understating. CE

[7] https://github.com/SZTAKI-EMI-PG-public/RS4SCPPS/tree/main/src.

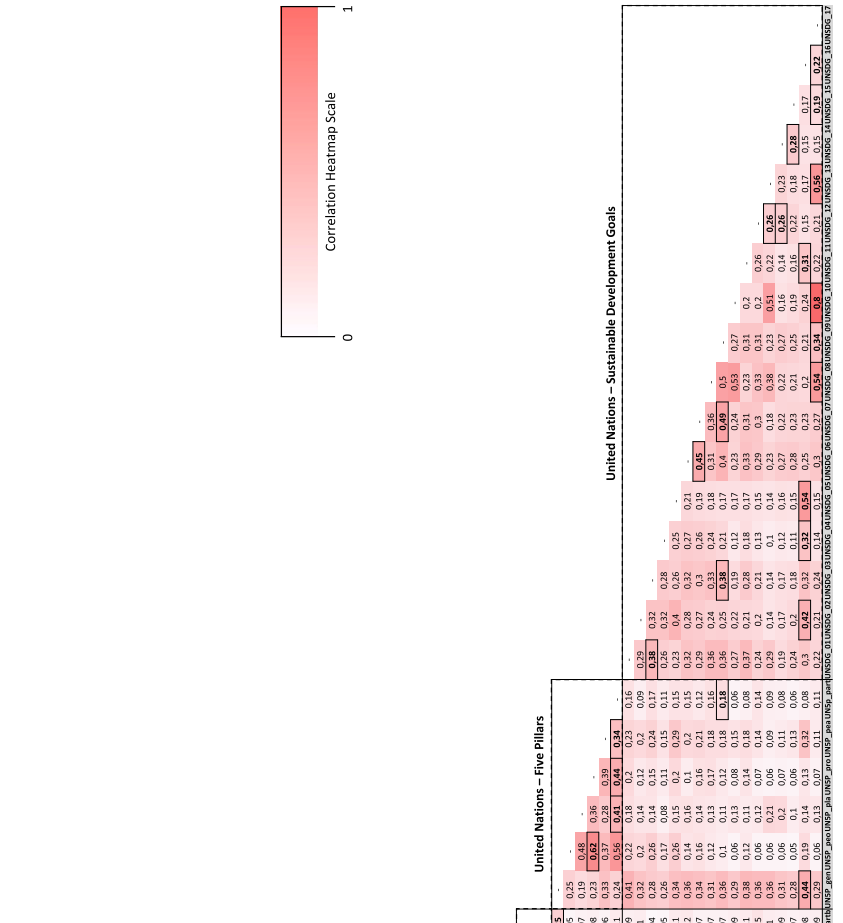


Fig. 2. Direct correlation matrix.

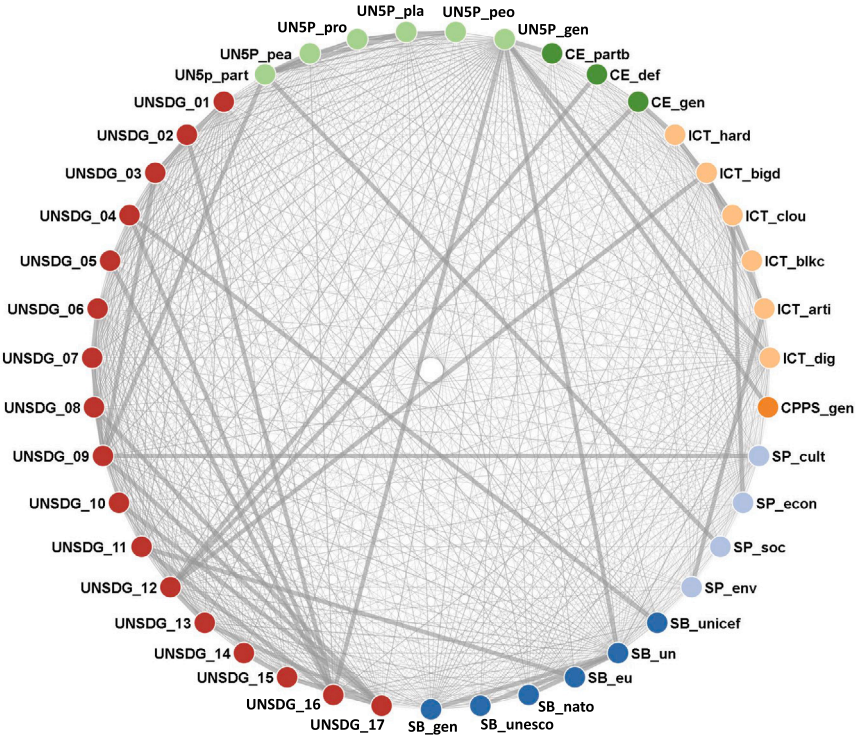


Fig. 3. DC network (highest correlation for a node is highlighted by a thicker gray chord).

still has to prove its competitiveness as a sustainable business alternative to the linear traditional production pattern (*take-make-use-dispose*), but the results of this investigation seem to confirm the (open) question of whether CE can actually be a solution towards sustainability, if involved actors do not even share a common portfolio of actionable terms.

4.1.2. Sustainability pillars

From the above analysis, two outcomes appear immediately evident. The first is that within their own domains, sustainability pillars show a very low level of correlation. Concepts like environment, society, economy and culture, which tend to be well-linked on the promotional front side of sustainability, loose this apparent cohesion when focusing on the secondary level of similarity. All DC scores are all below 0.15, with a maximum number of 19 common terms. On the other side, there seem to be stronger correlations between the pillars and ICT domain: AI (*ICT_arti*) reached the highest DC score of 0.32 in relation to environmental implications of sustainability (*SP_soc*), followed by the cloud technology (*ICT_clou*) with a 0.26 related to economic aspects. These results seem to confirm the existence of actual sustainability reasons behind the scepticism on the exploitation of these technologies of the future on a large scale. CE, despite its promising mission of a game-changer in sustainable business, shows again a relatively low level of correlations, which indicates the lack of a common vocabulary in addressing sustainability issues. It is worth highlighting how CE scored the highest DC in relation to the societal implications (*SP_soc*) of building sustainable partnerships (*CE_partb*). This is a fundamental element of the upstream and downstream (indirect) activities of a company's value and supply chain (so-called Scope 3). This aspect is also confirmed by the view of the UN, with a DC of 0.22 between *UN5P_part* and *SP_soc*. Another non-trivial result which is worth mentioning is the 0.35 between *SP_cult* and *UNSDG_09*, which indicates how the level of the industry, innovation and infrastructure in a society is highly correlated with its cultural identity. They are not only a mere technological affair, but rather influenced by the complex ecosystem of existential values and beliefs of a population.

4.1.3. Cyber-physical production systems

There is a confirmation of the preliminary assumption: CPPS domain seems to be quite distant from any other, in general (only ICT scored higher than 0.1). This might be due to the fact that scientific and technical forums usually adopt peculiar forms of hermetic, application-centred goals and expressiveness. However, the real reason might also rely on the fact that CPPS have not yet really (or adequately) addressed sustainability issues. At least not through actionable solutions, practical measures and best-practising use cases. However, the highest DC in CPPS domain was scored with UN 5 pillars general description (*UN5P_gen*), indicating that some basic convergence of concepts can be surely identified.

4.1.4. Information and communication technology

We generally find higher DC results within the domain, with the confirmation that Big Data (*ICT_bigd*) plays a fundamental role in conjunction with digitalization (*ICT_dig*) and AI (*ICT_arti*), whereas the latter shows an important link to blockchain technology

(*ICT_blkc*), as well. Another important result which is worth mentioning is that despite the voluptuous feeling produced by the term cloud (*ICT_clou*), there is a crucial influence and impact on sustainability by the hardware resources (*ICT_hard*) necessary to provide the virtualization of the architectural services (X-as-a-Service, shortly).

In general, all scopes in ICT domain show a coherent correlation and influence with each other, a symptom of the fact that many of these technologies are clearly interlaced. Targeting sustainability achievements and regulations in future society will most probably mean to have an impact on the utilization of all these instruments as a whole. It is hard to imagine, for example, AI (or blockchain) without an elementary process of digitalization/virtualization of resources and processes, big data and cloud. This will require (more than) adequate physical resources to support all the involved service chains. On a large, global scale, this inevitably and overwhelmingly affects the concept of *sustainability of ICT*, as currently conceived. This argumentation is also well supported by the non-trivial correlation emerged between ICT hardware scope and the general conceptualization of the CE (*ICT_hard vs. CE_gen*).

Digitalization, in general, seems to indicate (or confirm) a sort of evolutionary path for the sustainability of modern societies, as also prescribed by the UN 5 pillars (*ICT_dig* and *UN5P_gen*). But digitalization will need to coherently comply with guidelines also deriving from the other domains and scopes, in order to be truly sustainable. In particular, digitalization will have to deal with most-recurring correlation concepts (terms) analyzed in next section.

4.1.5. Circular economy

The most relevant DC of CE with the other domains is reached between the scope of partnership building (*CE_partb*) and documents discussing the general position of UN on the 5 pillars (*UN5P_gen*). This result is partially confirmed also by the DC with the UN5P-specific scope (*UN5P_part*). Other two DC results are worth mentioning here: the 0.38 between *CE_gen* and *UNSDG_12*, and the 0.16 between *CE_def* and again *UNSDG_12*. In both cases, the message emerging from the correlation is that circular economy, in general, or through its possible definitions, will inevitably have to deal with a *more responsible consumption and production*.

4.1.6. United Nations - five pillars

Documents in this domain show a generally high level of DC. This is probably due to the fact that all scopes addressed are organized by the same regulator (UN), as it happens also for UNSDG publications. It is nevertheless interesting to report that, even at a *non-trivial* level of correlation, concepts like *people, peace, prosperity, partnership and planet* keep sharing common existential requirements. Moreover, the high DC between *UN5P_gen* and *UNSDG_16* is worth separate mentioning, as it conveys the fundamental view that sustainability is heavily connected with *peace, justice and strong institutions* of a “healthy” society.

4.1.7. United Nations - sustainable development goals

All UN SDGs share a very high level of DC within their domain, even at this second level of terminology distance. This is explainable by the fact that authors of contents belong to same organization and implicitly share a common visions, and accordingly, common glossary and communication intentions. It is very interesting to highlight, however, that the highest value of DC within the UNSDGs was obtained between the *UNSDG_10* and the *UNSDG_17* documents, which directly relate the *reduction of inequalities* in the world with the search of the opportune *partnerships for goals*. In other words, the creation of alliances with partners should embrace and share a common vision of sustainable development, to be back-boned by an indispensable win-win scenario for profitability and possibility. Another very interesting but less surprising DC results emerged between the following cases: gender equality (*UNSDG_05*) and the purposes of peace, justice and strong institution (*UNSDG_16*); direct work and economic growth (*UNSDG_08*) with partnerships for goals (*UNSDG_17*); and, very interestingly, the reduction of inequalities (*UNSDG_10*) with the climate actions (*UNSDG_13*).

Outside the UNSDG domain, other results are worth mentioning, as well. For instance the 0.43 score between *UNSDG_04* and *SB_unicef*, which indicates how important the *quality of education* of new generations is for sustainability and how this will crucially impact the achievement of the overall goals. This is confirmed by reflection from the EU, which seems to be concentrating most of its efforts on *sustainable cities and communities*, based on the 0.25 score between *SB_eu* and *UNSDG_11*.

The gap between CPPS and UNSDG domains remains clear, with the most noteworthy result being that achieved between *CPPS_gen* and *UNSDG_06*, which relates CPPS in general with *clean water and sanitation* in the environment.

4.2. Term dimension

Outcomes of document correlations can be elaborated and also presented from the point of view of the correlation terms occurrences and their relative weights, as introduced in section 3.7. This gives a very peculiar footprint of the document connections by providing the actual (hidden) conceptual links between them.

The complete list of correlation terms count several hundreds of terms. We have extracted a list of the first 40 ones (based on the occurrence ranking), as reported in Fig. 4 (the illustration visually emphasises the scores reported in the first column of the table). We will not elaborate on all the terms of this list, but rather focus on the implications of those ones which are more relevant for our application domain and expertise related to CPPS. It is worth reminding that the core of our methodology is based on text analytics techniques and not on semantics analysis: any speculation on the “real” phrasal connection or order of terms in hypothetical sentences are left to the reader’s interpretation. One fact remains certain: the occurrences of a term in DCs is the symptom of its importance in the targeted corpus and forums, and as such, it should be adequately addressed as main driver in sustainability discussions.

Term	No. of DCs containing it	Total weight in DCs	Weight/DC Rate
include	**666**	112502	169
global	**528**	223859	424
service	**465**	85662	184
policy	**406**	64733	159
access	**351**	85446	243
impact	**300**	52006	173
change	**276**	71794	260
target	**276**	73217	265
support	**253**	34788	138
resource	**253**	55470	219
population	**210**	59802	285
share	**210**	38471	183
international	**171**	26658	156
national	**153**	53418	349
human	**153**	29894	195
source	**153**	23522	154
public	**153**	32731	214
risk	**136**	22265	164
age	**120**	49502	413
process	**120**	15770	131
pandemic	**105**	15135	144
asia	**105**	12908	123
africa	**105**	17034	162
continue	**105**	13470	128
measure	**105**	16166	154
capacity	**105**	10151	97
covid	**91**	8877	98
financial	**91**	14038	154
sector	**91**	17253	190
region	**91**	18174	200
practice	**78**	13702	176
protection	**78**	17444	224
government	**78**	13452	172
programme	**78**	10003	128
challenge	**78**	8993	115
integrate	**66**	8271	125
live	**66**	16069	243
territory	**66**	10484	159
plan	**66**	13084	198
person	**66**	9599	145
...	...	...	...

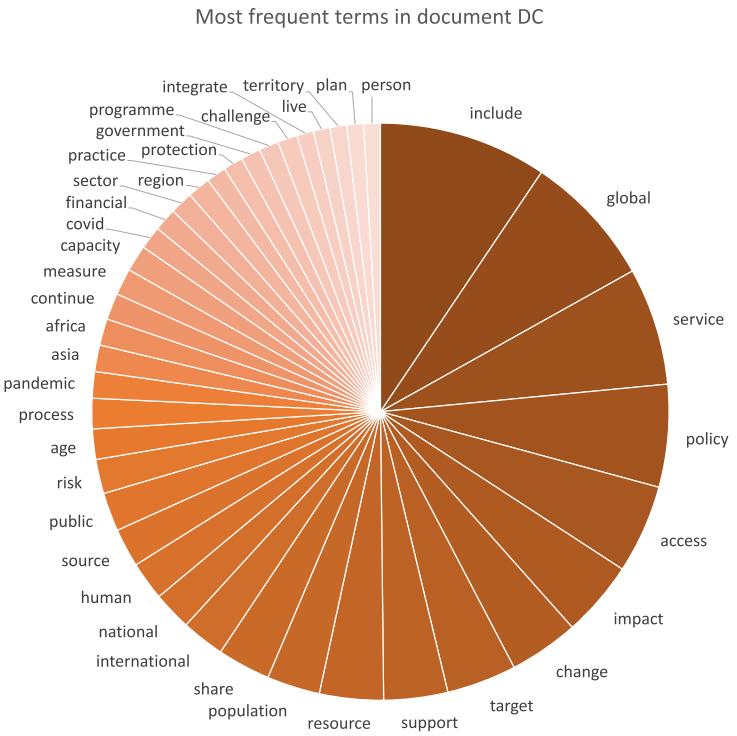


Fig. 4. First 40 DC terms and their occurrence ranking (complete list on Git repository).

4.2.1. Occurrence based ranking analysis

This section introduces the DC term-ranking results, a brief list and analysis of all major terms, as well as an explanation of the aggregated ranking dimension reported on Fig. 5.

Inclusion (666): the undisputed winner. The most occurring term in all correlations. The quantitative proof of EU's motto, as to which "*nobody has to be left behind*". Inclusion and diversity seem to represent fundamental assets to a company's sustainability efforts. They are indicated as an improvement to productivity and efficiency (humans are the most critical and fundamental resource), a support to better decision-making, and a means to make it easier to adopt structural and operational changes that benefit people, the planet, and their bottom. Inclusion can develop genuine connections between environmental sustainability and social justice, a concept that challenges traditional sustainability practices, which focus almost exclusively on the environmental dimension and ignore or overlook important social justice issues.

Service (465): services, with their importance in the global economy alongside manufacturing, are vital in a country's economic growth. Due to the world's shift towards global production, service-based society is likely to transform not only the composition of the world's economic production and employment, but potentially global trading patterns, too. And CPPS are no exception to this paradigm shift. This vision of the service economy indicates that the specialization in high-added-value manufacturing services is one of the fundamental concepts that might ensure the realization of sustainable economic growth and satisfy market requirements.

Access (351): services and resources need to be accessed in order to be profitable. This can not be seen or said in another way. Any obstacle or absence of access to the right information, service, or resource, at the right time, is a crucial barrier to any stability and sustainability of an organization and civil society.

Change (276): traditional approaches to production (or growth in general) have shown their limitations, mainly if we consider the linear model of the manufacturing industry. It has been argued that novel business paradigms (like CE, for example) still have to prove their actual beneficial effects in the long term and on a large scale. There is an undoubted concept widely returning in the corpus: society (companies, manufacturers, institutions, and so forth) needs to produce a change in the way objectives are currently pursued and achieved.

Resource (253): this is one of the most interesting correlation concepts of our investigations, together with service, source, share, process and capacity. Regardless of the domain and scope of documents analyzed, there seems to be an implicit convergence of views towards a fundamental aspect of our society and the earth: the capacity of our environmental resources (comprising humans) is limited and therefore, there is an impending need to provide collaborative production processes and services for the optimization of resource exploitation through capacity sharing.

Share (210): another pillar of our investigation outcomes. It is highly connected to the paradigm shift necessary to move from a traditional society of owning products and assets to a newer one of resources and capacity sharing. Many aspects and initiatives

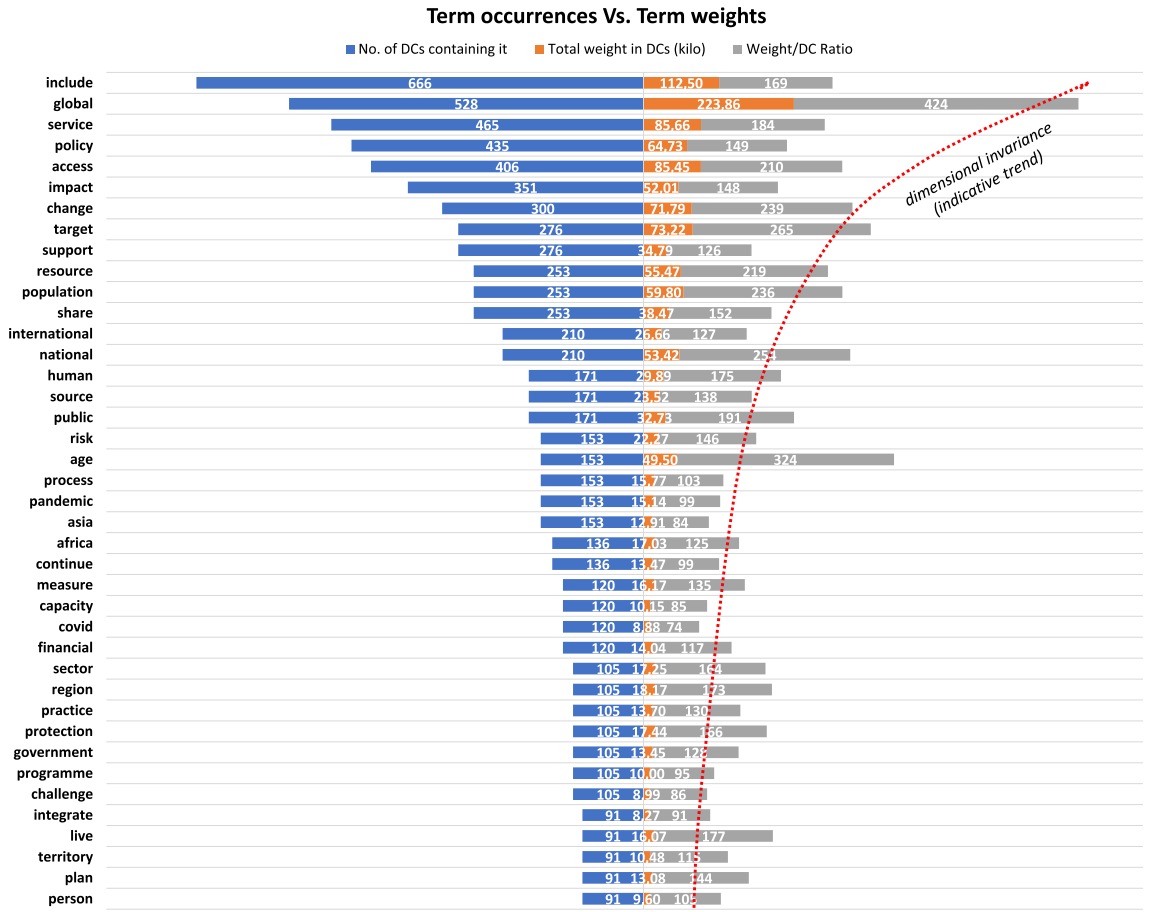


Fig. 5. Aggregation of DC ranking dimensions (term presence in DC is the leading one).

of the sharing economy (as a possible strategy in the broader CE, too) are still under investigation, but this result already shows how unconsciously important this concept is for the various regulators and opinion leaders from various domains. From a more strictly manufacturing point of view, sharing mechanisms for capacity and resources can possibly palliate the effects of unpredictable fluctuations in market demands, mainly for small and medium sized enterprises (SMEs), which can not usually afford impulsive restructuring investments.

Human (153): from any point of view of a possible interpretation (production, society, economy, environment), humans (and their age) are the most important and yet the most crucial and sensitive resource of the planet. Any product manufactured or any service provided in modern society is intended for humans. There can not be any plan or strategy for the planet of the future without taking into account the human factor, which is fundamental for the social and economic dimension of sustainability.

Source (153): this can be interpreted as the root word for re-source, and as such, its score further reinforces the relevance and the implications of the latter. Nevertheless, the peculiarity of this word could also indicate the connection to water supply, the support for biodiversity, the creation of natural habitats, the maintenance of wetlands and in general, all those conditions closely linked to geodiversity and the sustainable use of resources.

Process (120): any organic activity is seen as a process, and driving sustainability means achieving the necessary processes (and their optimization). According to the World Economic Forum, sustainability is the new digital and is disrupting the economy just as digital did. Whatever domain we consider, sustainability will be at the heart of all processes. On the other hand, only adequate transformation activities, in terms of input/output, can guarantee the necessary sustainability structure.

Measure (105): from an engineering point of view, one can only improve what can be measured. In 1883, Lord Kelvin stated: "*If you cannot measure it, you cannot improve it*". This is still true, and it is for any of the terms already seen before (resource, access, process, inclusion, and so forth). Without a measure, for example, of the current carbon footprint of manufacturing processes, no improvement can be put into place. Sustainability plans start with measuring the current situation (as-is) of our (business) activities/products and deriving the necessary improvement patterns/pathways for sustainable growth.

Capacity (105): in this bare form, it might be susceptible to different interpretations. In any case, the importance of capacity is that it allows a business to fulfil demand levels and provide flexibility to make diverse products. Capacity is the highest total level of output that can be produced by a company in a certain time period. Efficient capacity management ensures that no extra money

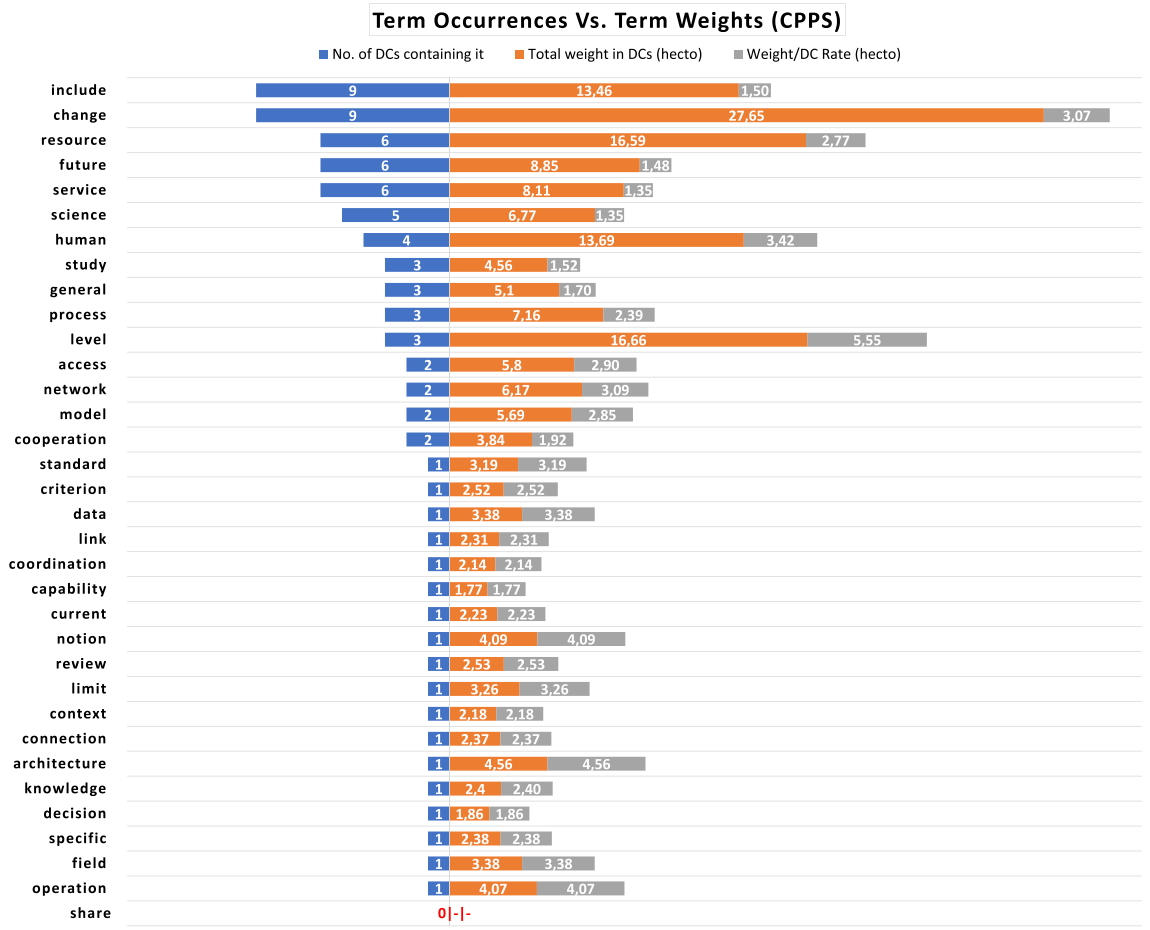


Fig. 6. DC aggregated ranking results for CPPS.

is spent. This means, in turn, that resources are optimized for the achievement of the same purpose. This is an essential concept towards sustainable manufacturing.

Integrate (66): this word is highly related to the top-rated one in the list, that is *inclusion*. It represents its natural counterpart when referring to resources (even humans) in any application environment, from institutional to manufacturing to information technology. Integration can furthermore trigger standards for interoperability and communication between elements operating on different platforms. From a societal perspective, it is the intermixing of entities that were previously isolated. Integration is the driver for complex service provision, resource sharing and diversification of company offers.

4.2.2. CPPS specific outcome

Illustration in Fig. 6 highlights the results of the analysis performed specifically in the CPPS domain. The most striking result is the complete absence of the term *share* from all CPPS DCs. This might be viewed as a confirmation of the existence of a conceptual gap or scepticism, or practical barriers in relation to this production and business mechanism in CPPS. Nevertheless, the presence of terms such as *include*, *access*, *cooperation* and *connection*, might indicate the existence of a *masked* correlation with the concept of *sharing* (future works, see Sect. 5). On the other hand, many terms investigated in the previous section (change, resource, service, and so forth) are also confirmed in CPPS-specific outcomes, highlighting how CPPS are to be seen not as an isolated and futuristic reality but as integrating elements of the entire society. It is worth emphasizing how the strength of the connections of CPPS with other domains are generally low, as also illustrated by cords in Fig. 3. Moreover, the aggregated ranking reported in Fig. 6 evidences how higher DC occurrences (from 4.0) relate CPPS to other interdisciplinary domains, with *human resource* (again) at the centre of the global *change* for the *future*. This was one of the criticisms addressing Industry 4.0 and the dilemma of finding the right balance between fully automated and (human) collaboration-enhanced CPPS.

As regards specifically resource sharing, other research works like [58] already indicate it as a viable support to strategic collaborations in federated companies (SMEs), helping them optimize the utilization of their assets, avoid shortages and minimize the environmental impact of their logistics. Leveraging the same extension, also CPPS can be seen as a federated collaboration entity, with behaviour, strategies and issues typical of a traditional company. Note that such a collaboration can be supported by the recently proliferating methods and business models of *platform-based manufacturing* [59].

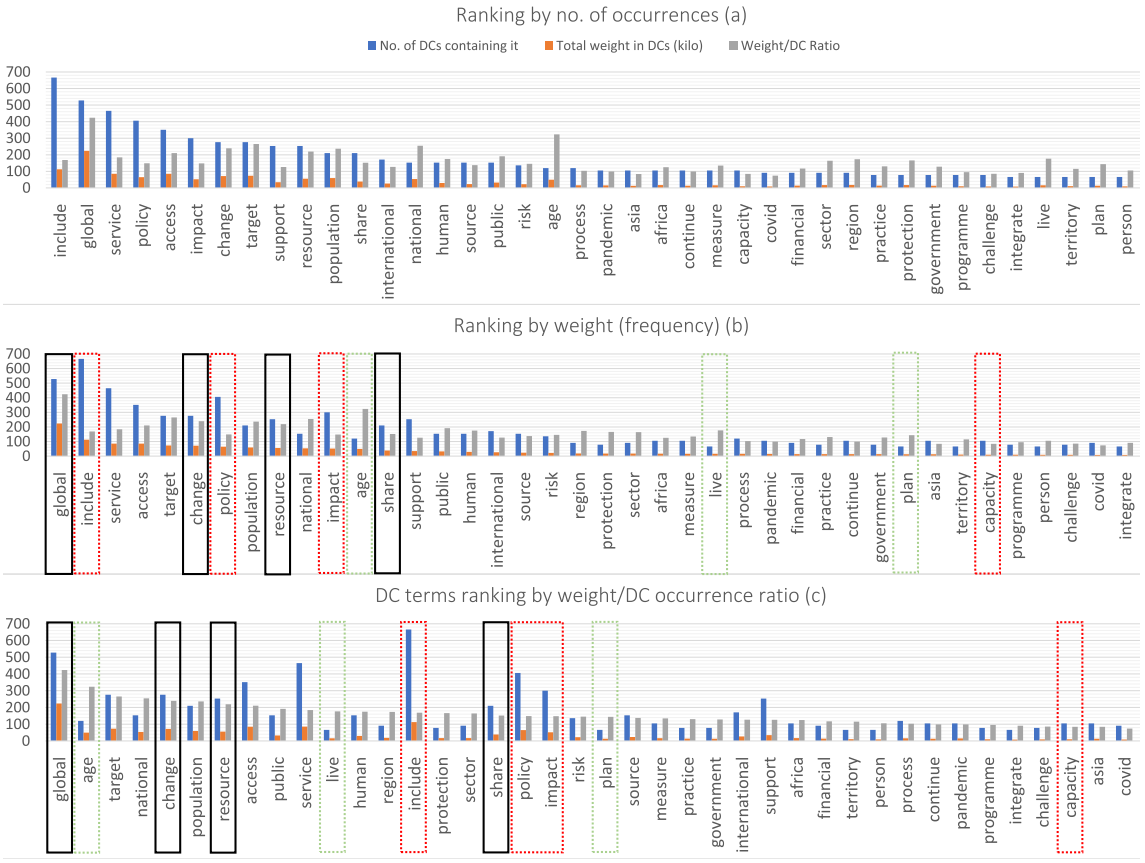


Fig. 7. Comparison of DC terms ranking dimension.

4.2.3. Ranking dimension comparison

An interesting outcome of our investigations concerns the different positioning of a term in the three DC ranking dimensions: i) term occurrences, ii) total and iii) average weight. Reorganizing results depicted in Fig. 5 (order is driven by the term occurrences – the blue component of the aggregated item lane) and reordering them according to the specific dimensional ranking, we obtain the comparative illustration visualized in Fig. 7. The aim is to see to what extent the ranking of terms is susceptible to variation in relation to the three correlation dimensions and whether there are *stable* indicators emerging from the comparison. The transition from the first dimension (DC term-occurrence) to the other two can produce three different ranking-variation scenarios: i) a better positioning, ii) a worse one, and iii) (almost) an invariance in the positioning of the term.

In the first case, we have selected the following exemplary DC terms: *include, policy, impact* and *capacity* (highlighted in a red dotted rectangle in Fig. 7). As appears from the illustration, these terms lose ranking positions when analyzing the evolution from (a) to (c) of the illustration. This is due to the fact that there are considerable differences in term frequencies along with the corpus (higher dispersion), and it is also confirmed by the average DC weight of these terms. This can also be noticed visually from the aggregated representation in Fig. 5, which should be (horizontally) symmetric and funnel-shaped. On the contrary, it is quite fragmented and unordered on the right side.

For the second case, we could take into account the following words: *age, live* and *plan*. As opposed to the previous scenario, these terms gain ranking positions with dimensional traversing. This is explainable by the fact that even though they occur in fewer DCs in the corpus, when they do, they have a much heavier relevance (weight) for the involved documents. Like in the previous case, these terms have an aggregated representation diverging from the expected one.

As regards the third case, from Fig. 7 a possible message can be as follows: *global change and resource sharing*. These terms represent a sort of invariant pattern in all of the three DC ranking dimensions. In b) and c) sub-figures, their distribution results are (quite) uniform if compared to the number of occurrences they have and the total weight calculated in the same DC matching pair-set. In other words, there is a constant balance between documents of the corpus, which project these terms in the core of sustainability concepts. The indicative boundary of such a dimensional invariance was derived by simple interpolation of values and is depicted on the right side of Fig. 5 (red-dotted line). It is worth highlighting once more that this research work focused on the identification of document correlation terms and not correlation between correlation terms. This is undoubtedly a further interesting dimension of investigation, and some preliminary outcome has already been outlined in the paper. Correlation term ranking illustrated in Fig. 5 (whole corpus) and 6 (CPPS dimension) implicitly indicate an order of reading for such terms.

5. Validity and limitations

Our research work is not (yet) another method for literature review, but an almost fully automated methodology for unrevealing hidden correlations in (publicly) available corpus, in the light of HITECS conceptualization. The dimension of the investigated corpus aims at the purpose of HITECS and guarantees an acceptable compromise between the minimum coverage of the domains and the manual effort necessary in the pre-processing phase of the methodology. In compliance with the prerequisites of HITECS, the corpus is still sufficient to produce evidence of possible, hidden correlations. All correlation terms gain unbiased relevance, being the result of automated procedure and converging to the concept of the raking invariance, as introduced in the paper, as well. HITECS seeks non-trivial correlations between documents, not between terms. This sort of correlation analysis is out-of-scope in this paper but represents indeed an interesting dimension of further investigation. The paper brings this aspect to light indirectly: some preliminary outcomes can be already outlined in Figs. 5 and 6, where the ranking of correlation terms also indicates a possible order for reading such terms, being the occurrences of a term in DC a metrics also of the probability distribution of such a term across the entire corpus.

Major goal of HITECS is to provide a way for seeking concrete and actionable aspects of a concept (in our case we focused on "sustainability"), at any level, in any domain. These concepts are disclosed starting from HITECS findings (hidden terms correlated by eliminating first-level key-words). Eliminating the fuzziness around the meaning of an inflexed term driver (slogan) can effectively help the decision-making process of a business, avoiding ineffective or useless investments due to main-stream influence and/or overuse. What really counts in HITECS is the terminology used in background, distributed, or concentrated, around a main concept of investigation. In HITECS, disjoint concepts can come to the spotlight and open new pathway of investigation or research. What really matters is that HITECS provide a tool to capture hidden terms to reason on in the light of a new sustainability strategy definition.

There are also some conceptional and operative constraints in the use of HITECS, as summarized hereafter. Sources collected from URLs consider only the text contained in the sanitized HTML: no mechanism is currently included in the methodology to recursively navigate the URL structure of a website in the search of text from separate documents attached to it. The corpus can be built from a combination of URLs and off-line documents, but the latter have to be in the preferred form of simple-text files, according to the third-party tool utilized in the HITECH pre-processing phase. Moreover, the tool does not provide at the moment an application programming interface (API) or software development kit (SDK) mechanism; the project has to be downloaded and run locally. HITECS focuses only on *direct correlation* analysis (exact coincidence of bare terms), *indirect* (synonyms and is-a relation), as well as *masked* correlations (existence of continuous, non-null correlation paths between documents) is part of the next works.

6. Conclusions and future work

New concrete production strategies need to be realised in a truly sustainable society of the future. Sharing resources has gained interest as strategy for sustainability, but only a few attempts have been made to assess its effectiveness and applicability, in particular in the case of CPPS. The reason for this is that the idea of sustainability necessitates disambiguation, as frequently used with different meanings in main-stream communication. There is an unrevealed connection between a concept and all the words used to convey it. To find these hidden relations, we presented HITECS, a novel, corpus independent text analytics methodology based on the combination of bare term frequency and cosine similarity, and whose basic strategy is the exclusion of all the trivial, first-level query-driving key-word matching from document correlations. As mentioned in the introduction, all master data, parameters and sources utilized in HITECS analysis are made publicly accessible via the GitHub repository reported in the section.

HITECS outcomes have spotlighted how, despite a current scepticism and a plethora of opinions around sustainability, scientific and societal discussions are naturally evolving towards hidden direction indicated by terms such as *include, global, access, service, integration, resource, and share*, among the others, being implicitly correlated to a sustainable development. These concepts are much wider than the sole environmental footprint, strengthening the primordial vision which considers the planet and its environmental resources, the global population and its societal and economic aspects as a whole sharing system. Sustainability is not only an environmental issue and sustainable development can surely benefit from a balanced and well-fared societal outfit.

CPPSs, being them the core part of future digitized, productive societies, cannot redeem from being aligned with this vision of responsible and sustainable production. Developing sustainable CPPSs will inevitably mean to explore the implications of the emerged correlation terms in a new vision of manufacturing and organizational practices, where resource access, integration and sharing can be a fundamental asset for societal stability, the unmissable prerequisite for any form of development. At the moment there is little evidence that CPPS domain is giving the proper importance to such concepts. Nevertheless, resource sharing appears to be a promising confirmation and an actionable practice for enabling sustainability in CPPS, too. In future investigations, corpora will be extended with recent policy materials addressing *security* and *equity* concerns, as these two dimensions, together with sustainability (a so-called SES framework) are the key elements of a transformative change. The method proposed in this paper can help find and focus on key underlying common concepts and their relations even though the SES framing is notoriously ambiguous, both across disciplines and policy sectors [60]. Nevertheless, future works will give space to hidden, indirect correlation (synonyms and is-a terms) and masked correlation (existence of continuous, non-null correlation paths) in documents, further enhancing the effectiveness and usefulness of HITECS in analysing what is not evident in a text.

CRediT authorship contribution statement

Gianfranco Pedone: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Formal analysis, Data curation, Conceptualization. **József Váncza:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Ádám Szaller:** Writing – original draft, Conceptualization.

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.

Data availability

Data associated with our study been deposited on a publicly available repository, as also referenced in the article: github.com/SZTAKI-EMI-PG-public/RS4SCPPS.

Acknowledgement

The authors thank for the support of the NRDIO ED-18-2-2018-0006 grant on “Research on prime exploitation of the potential provided by the industrial digitalization” and of the TKP2021-NKTA-01 grant on “Research on cooperative production and logistics systems to support a competitive and sustainable economy”.

Appendix A. Methodology details

This section reports additional conceptual and computational details about the methodology steps presented in section 3.

A.1. Document pre-processing

When documents were openly and directly processable (like open access publications or open organizational repositories and URLs), this functionality was directly provided by the (external) service vectorization. On the other hand, there were cases in which documents had to be downloaded, first, in the form of offline textual manuscripts and then provided to the vectorization service. In these cases, the sanitization procedure was manually executed by the authors of this paper. This was a simplification in terms of programming efforts, as the entire process could be automated by programming-specific web crawlers (out of scope here). Excluded first-level keywords represented the querying logic for the corpus identification.

A.2. Document vectorization

In the extreme case of all disjoint documents (unique terms), we will have a vocabulary of *N*100* words, where N is the number of documents in the corpus. In case the term does not exist in a particular document, that particular TF value (weight) will be 0 for that particular document. In our investigations, the number of columns in (1) will range between 100 (all the extracted terms are in common in the entire corpus) and 4300 (extreme case where there is no term in common between two arbitrary documents), inclusive. In the latter case, we can already predict that the correlation matrix will be made of zeroes except diagonal elements (2). Moreover, where a term vector does not reach the 100 items, we add an invariant place-holder term with weight 0 (zero), until all the 100 positions are complete (ex. “t1 0”, “t2 0”, etc.).

For grammatical reasons, documents are going to use different forms of a word, like, for example *build, builds, and building*. Additionally, there are families of derivative related words with similar meanings, such as *democracy, democratic, and democratization*. In practical cases, a text analytics technique searching for one of these words will also return documents that contain another word in the set. Fundamental in vectorization is lemmatization. Lemmatization usually leverages the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*. Stemming is another technique which aims at reducing the inflectional forms of terms but usually refers to a crude heuristic process that truncates the ends of words in the hope of achieving this goal correctly most of the time and often includes the removal of derivational affixes.

A.3. Document normalization

Recalling $w_{i,j}$ from (1) (which is the weight of the j-th term in n-th document), then we can define the weight of a given i-th document as the sum of all its term-vector weights (frequencies):

$$DW_i = \sum_{j=1}^{m} \mathbf{w}_{i,j} \tag{A.1}$$

where $m = 100$ for each specific document j.

At this point, the heaviest document in a specific domain scope will be:

$$HD_{scope} = \max\{DW_j\} \tag{A.2}$$

where W_i is the i-th row in (1) and $j \in$ set of indexes in the specific scope (*local/global*).

We can now calculate the term-vector scaling rate (TSR) for a given document in a given scope as:

$$TSR_{scope,i} = \frac{HD_{scope}}{DW_i} \tag{A.3}$$

and finally, the scaled term-vector as:

$$SW_{scope,i} = W_i * TSR_{scope,i} \tag{A.4}$$

The goal is to have only 1 term-vector representing a specific domain scope in the end. We initially had 51 documents. After the local normalization, this number was reduced to 43. The number of the selected doc is arbitrary; this initial corpus was sufficient for the overall investigation purposes.

Global normalization (corpus), on the contrary, consisted of the application of linear scaling to all the documents present in the corpus. After this process, all documents in the corpus will have the same weight (and so they are of the same order of textual relevance): only the terms and their frequency are the actual discriminants for correlation, as to be.

A.4. Cosine similarity

We might say that CS learns the context according to the formula below:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{l} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{l} (\mathbf{x}_i)^2}\sqrt{\sum_{i=1}^{l} (\mathbf{y}_i)^2}} \tag{A.5}$$

where

$$l \leqslant n * m,$$

$n =$ number of documents in the corpus, and $m =$ number of extracted terms per document.

CS ranges between zero (0) and one (1), inclusive. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no *conceptual* match. The closer is the cosine value to 1, the smaller is the angle and therefore, the greater is the match between the two vectors. CS will actually indicate the level of correlation (conceptual distance) between documents.

A.5. Correlation analysis

The outcome of similarity computation between two documents can be formalized in terms of cosine similarity values and correlation terms, as follows:

$$doc_i | doc_j : c_{i,j}(cterm_1)(cterm_2)\dots(cterm_k) \tag{A.6}$$

where i, j are indexes from the normalized corpus, $c_{i,j}$ is the correlation value between *i*-th and *j*-th document, whilst $(cterm_k)$ is the *k*-th common term found in the relationship. As the max value for *k* is 100 (total match between the two term-vectors in terms and weights), the cardinality of the common term set practically indicates a *percentage*, too (which is not equal to, and not to be confused with $c_{i,j}$).

Major computational aspects are reported hereafter:

i) DC term occurrences ($dcterm_{occur}$) have been calculated simply by iterating over all the document DC term-pairs and scoring a +1 when the term is present (note that DC is a symmetric relation. Therefore iterations are calculated only for one pair of documents, on cascade, for simplicity from the first to the last document in the corpus - Table 2);
ii) As introduced in cosine similarity (section 3.6), correlation scores whenever there is a term-pair matching between two document term-vectors. The term has a not null value (weight) on the same t_m column for the two documents, as reported in (1).

We defined $dcterm_{totWeight}$ as the sum of all the term's weights appearing in any correlation term pairs containing it.

Note: the weight of a term has been considered only *once* in all of the correlation term-pairs related to a specific document. This was necessary in order to maintain correlation consistency, as a DC term will inevitably repeat in all the matching term pairs for a given document.

Explanation: similarity distance is symmetric and calculated on cascade (the order of documents is *irrelevant*). For this reason, the weight of a given matching term need to be considered only once in all of the DC term pairs produced for a given document.

iii) The average weight of a term in correlations is defined as (rounded to the closest integer number for simplicity):

$$dcterm_{avgWeight} = dcterm_{totWeight} / dcterm_{occur} \tag{A.7}$$

References

[1] L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, K. Ueda, Cyber-physical systems in manufacturing, CIRP Ann. 65 (2) (2016) 621–641, https://doi.org/10.1016/j.cirp.2016.06.005.

[2] T. Kaihara, N. Nishino, K. Ueda, M. Tseng, J. Váncza, P. Schönsleben, R. Teti, T. Takenaka, Value creation in production: reconsideration from interdisciplinary approaches, CIRP Ann. – Manuf. Technol. 67 (2) (2018) 791–813, https://doi.org/10.1016/j.cirp.2018.05.002.

[3] J. Váncza, L. Monostori, D. Lutters, S. Kumara, M. Tseng, P. Valckenaers, H. Van Brussel, Cooperative and responsive manufacturing enterprises, CIRP Ann. – Manuf. Technol. 60 (2) (2011) 797–820, https://doi.org/10.1016/j.cirp.2011.05.009.

[4] M. Radetzky, L. Grams, B. Ulutas, S. Bracke, Sustainability versus efficiency of manufacturing process: structured comparison of two high precision fine grinding processes, in: 25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing August 9–14, 2019, Chicago, Illinois, USA, in: Procedia Manufacturing, vol. 39, 2019, pp. 859–867.

[5] T. Bosschaert, Circularity is not sustainability: how well-intentioned concepts distract us from our true goals, and how SiD can help navigate that challenge, in: The Impossibilities of the Circular Economy, Routledge, 2022, pp. 72–80.

[6] T.H. Mason, C.R. Pollard, D. Chimalakonda, A.M. Guerrero, C. Kerr-Smith, S.A. Milheiras, M. Roberts, P.R. Ngafack, N. Bunnefeld, Wicked conflict: using wicked problem thinking for holistic management of conservation conflict, Conserv. Lett. 11 (6) (2018) e12460, https://doi.org/10.1111/conl.12460.

[7] G. Caniglia, R. Freeth, C. Luederitz, J. Leventon, S.P. West, B. John, D. Peukert, D. Lang, H. von Wehrden, B. Martín-López, et al., Practical wisdom and virtue ethics for knowledge co-production in sustainability science, Nat. Sustain. 6 (2023) 1–9, https://doi.org/10.1038/s41893-022-01040-1.

[8] A.A. Alazmi, H.S. Alazmi, Closing the gap between research and policy-making to better enable effective educational practice: a proposed framework, Educ. Res. Policy Pract. 22 (2022) 91–116, https://doi.org/10.1007/s10671-022-09321-4.

[9] J. Kirchherr, D. Reike, M. Hekkert, Conceptualizing the circular economy: an analysis of 114 definitions, Resour. Conserv. Recycl. 127 (2017) 221–232, https://doi.org/10.1016/j.resconrec.2017.09.005.

[10] J. Kirchherr, N.-H.N. Yang, F. Schulze-Spüntrup, M.J. Heerink, K. Hartley, Conceptualizing the circular economy (revisited): an analysis of 221 definitions, Resour. Conserv. Recycl. 194 (2023) 107001, https://doi.org/10.1016/j.resconrec.2023.107001.

[11] G. Pedone, R. Beregi, K.B. Kis, M. Colledani, Enabling cross-sectorial, circular economy transition in sme via digital platform integrated operational services, Proc. Manuf. 54 (2021) 70–75, https://doi.org/10.1016/j.promfg.2021.07.048.

[12] A.-H. Tan, H. Mui, K. Terrace, Text mining: the state of the art and the challenges, in: Proceedings of the Pakdd 1999 Workshop on Knowledge Discovery from Advanced Databases, vol. 8, 1999, pp. 65–70.

[13] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimed. Tools Appl. (Jul 2022), https://doi.org/10.1007/s11042-022-13428-4.

[14] S. Zhang, D. Zhu, Have countries moved towards sustainable development or not? Definition, criteria, indicators and empirical analysis, J. Clean. Prod. 267 (2020) 121929, https://doi.org/10.1016/j.jclepro.2020.121929.

[15] H. Zeng, X. Chen, X. Xiao, Z. Zhou, Institutional pressures, sustainable supply chain management, and circular economy capability: empirical evidence from Chinese eco-industrial park firms, J. Clean. Prod. 155 (2017) 54–65, https://doi.org/10.1016/j.jclepro.2016.10.093.

[16] M.P. Pieroni, T.C. McAloone, D.C. Pigosso, Business model innovation for circular economy and sustainability: a review of approaches, J. Clean. Prod. 215 (2019) 198–216, https://doi.org/10.1016/j.jclepro.2019.01.036.

[17] A.C. Silvério, J. Ferreira, P.O. Fernandes, M. Dabić, How does circular economy work in industry? Strategies, opportunities, and trends in scholarly literature, J. Clean. Prod. 412 (2023) 137312, https://doi.org/10.1016/j.jclepro.2023.137312.

[18] A. Laurent, C. Molin, M. Owsianiak, P. Fantke, W. Dewulf, C. Herrmann, S. Kara, M. Hauschild, The role of life cycle engineering (lce) in meeting the sustainable development goals–report from a consultation of lce experts, J. Clean. Prod. 230 (2019) 378–382, https://doi.org/10.1016/j.jclepro.2019.05.129.

[19] A. Hannon, E.G. Callaghan, Definitions and organizational practice of sustainability in the for-profit sector of Nova Scotia, J. Clean. Prod. 19 (8) (2011) 877–884, https://doi.org/10.1016/j.jclepro.2010.11.003.

[20] P. Roman, G. Thiry, C. Muylaert, C. Ruwet, K. Maréchal, Defining and identifying strongly sustainable product-service systems (sspss), J. Clean. Prod. 391 (2023) 136295, https://doi.org/10.1016/j.jclepro.2023.136295.

[21] J. Kantola, Y. Liu, P. Peura, T. de Leeuw, Y. Zhang, M. Naaranoja, A. Segev, D. Huisingh, Innovative products and services for sustainable societal development: current reality, future potential and challenges, J. Clean. Prod. 162 (2017) S1–S10, https://doi.org/10.1016/j.jclepro.2017.07.091.

[22] G. Beier, A. Ullrich, S. Niehoff, M. Reißig, M. Habich, Industry 4.0: how it is defined from a sociotechnical perspective and how much sustainability it includes – a literature review, J. Clean. Prod. 259 (2020) 120856, https://doi.org/10.1016/j.jclepro.2020.120856.

[23] W. Leal Filho, A.L. Salvia, R. Ulluwishewa, I.R. Abubakar, M. Mifsud, T.J. LeVasseur, V. Correia, A. Consorte-McCrea, A. Paço, B. Fritzen, S. Ray, N. Gordon, J.M. Luetz, B. Borsari, M. Venkatesan, S.A. Mukul, R.M. Carp, H. Begum, E.K. Nunoo, N. Muthu, S. Sivapalan, K. Cichos, E. Farrugia, Linking sustainability and spirituality: a preliminary assessment in pursuit of a sustainable and ethically correct world, J. Clean. Prod. 380 (2022) 135091, https://doi.org/10.1016/j.jclepro.2022.135091.

[24] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Inf. Softw. Technol. 55 (12) (2013) 2049–2075, https://doi.org/10.1016/j.infsof.2013.07.010.

[25] R. Raman, V.K. Nair, A. Shivdas, R. Bhukya, P. Viswanathan, N. Subramaniam, P. Nedungadi, Mapping sustainability reporting research with the un's sustainable development goal, Heliyon 9 (8) (2023) e18510, https://doi.org/10.1016/j.heliyon.2023.e18510.

[26] L. Irajifar, H. Chen, A. Lak, A. Sharifi, A. Cheshmehzangi, The nexus between digitalization and sustainability: a scientometrics analysis, Heliyon 9 (5) (2023) e15172, https://doi.org/10.1016/j.heliyon.2023.e15172.

[27] J. Han, M. Kamber, J. Pei, Getting to know your data, in: J. Han, M. Kamber, J. Pei (Eds.), Data Mining, third edition, in: The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston, 2012, pp. 39–82.

[28] UNDP, Human development index (HDI), https://hdr.undp.org/data-center/human-development-index#/indicies/HDI. (Accessed 19 March 2024), 2020.

[29] Youmatter, Sustainable development - what is it? Definition, history, evolution, importance and examples, https://youmatter.world/en/definition/definitions-sustainable-development-sustainability/. (Accessed 19 March 2024), 2020.

[30] Wikipedia, Sustainable development, https://en.wikipedia.org/wiki/Sustainable_development. (Accessed 19 March 2024), 2022.

[31] UNESCO, Sustainable development - building peace in the minds of men and women, https://www.unesco.org/en/education/sustainable-development#. (Accessed 19 March 2024), 2022.

[32] NATO, Environment, climate change and security, https://www.nato.int/cps/en/natohq/topics_91048.htm. (Accessed 19 March 2024), 2022.

[33] EU, Europe Sustainable Development Report, https://s3.amazonaws.com/sustainabledevelopment.report/2021/Europe+Sustainable+Development+Report+2021.pdf. (Accessed 19 March 2024), 2021.

[34] UN, Support sustainable development and climate action, https://www.un.org/en/our-work/support-sustainable-development-and-climate-action. (Accessed 19 March 2024), 2019.

[35] UNICEF, UNICEF and the sustainable development goals, https://www.unicef.org/sdgs. (Accessed 19 March 2024), 2019.
[36] EPA, Sources of greenhouse gas emissions, https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions. (Accessed 19 March 2024), 2022.
[37] Center for Sustainable Systems, University of Michigan, Carbon Footprint Factsheet, https://css.umich.edu/sites/default/files/carbon%20footprint_css09-05_e2021.pdf. (Accessed 19 March 2024), 2021.
[38] S. Woodcraft, N. Bacon, L. Caistor-Arendar, T. Hackett, Design for social sustainability, http://www.social-life.co/media/files/DESIGN_FOR_SOCIAL_SUSTAINABILITY_3.pdf. (Accessed 19 March 2024), 2022.
[39] ESG-The Report, The g in esg, https://www.esgthereport.com/what-is-esg/the-g-in-esg/. (Accessed 19 March 2024), 2022.
[40] K. Nurse, Culture as the Fourth Pillar of Sustainable Development, Commonwealth Secretariat Malborough House, Pall Mall, London, UK, 2006.
[41] X. Wu, V. Goepp, A. Siadat, Cyber physical production systems: a review of design and implementation approaches, in: 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2019, pp. 1588–1592.
[42] O. Cardin, Classification of cyber-physical production systems applications: proposition of an analysis framework, Comput. Ind. 104 (2019) 11–21, https://doi.org/10.1016/j.compind.2018.10.002.
[43] L.A. Cruz Salazar, D. Ryashentseva, A. Lüder, B. Vogel-Heuser, Cyber-physical production systems architecture based on multi-agent's design pattern—comparison of selected approaches mapping four agent patterns, Int. J. Adv. Manuf. Technol. 105 (9) (2019) 4005–4034, https://doi.org/10.1007/s00170-019-03800-4.
[44] UN General Assembly, Road map for digital cooperation: implementation of the recommendations of the high-level panel on digital cooperation, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/102/51/PDF/N2010251.pdf. (Accessed 19 March 2024), 2020 (A/74/821).
[45] A. van Wynsberghe, Sustainable AI: AI for sustainability and the sustainability of AI, AI Ethics 1 (3) (2021) 213–218, https://doi.org/10.1007/s43681-021-00043-6.
[46] P. Li, P. Jiang, Enhanced agents in shared factory: enabling high-efficiency self-organization and sustainability of the shared manufacturing resources, J. Clean. Prod. 292 (2021) 126020, https://doi.org/10.1016/j.jclepro.2021.126020.
[47] L.J. Nieuwenhuis, M.L. Ehrenhard, L. Prause, The shift to cloud computing: the impact of disruptive technology on the enterprise software business ecosystem, Technol. Forecast. Soc. Change 129 (2018) 308–313, https://doi.org/10.1016/j.techfore.2017.09.037.
[48] C. Cappiello, P. Meliá, P. Plebani, Modeling CO2 emissions to reduce the environmental impact of cloud applications, in: Advanced Information Systems Engineering Workshops, Springer International Publishing, 2016, pp. 155–166.
[49] F. Lucivero, Big data, big waste? A reflection on the environmental sustainability of big data initiatives, Sci. Eng. Ethics 26 (2) (2020) 1009–1030, https://doi.org/10.1007/s11948-019-00171-7.
[50] University of Oxford, Environmental impact of it: desktops, laptops and screens, https://www.it.ox.ac.uk/article/environment-and-it. (Accessed 19 March 2024), 2022.
[51] B. Krumay, R.B. Wu, Measuring the environmental impact of ict hardware, Int. J. Sustain. Dev. Plan. 11 (2016) 1064–1076, https://doi.org/10.2495/SDP-V11-N6-1064-1076.
[52] S. Kara, M. Hauschild, J. Sutherland, T. McAloone, Closed-loop systems to circular economy: a pathway to environmental sustainability?, CIRP Ann. 71 (2022), https://doi.org/10.1016/j.cirp.2022.05.008.
[53] OECD, Successful partnerships - a guide, https://www.oecd.org/cfe/leed/36279186.pdf. (Accessed 19 March 2024), 2022.
[54] WHO, The 2030 agenda for sustainable development and the un decade of healthy ageing 2021-2030, https://unfoundation.org/blog/post/the-sustainable-development-goals-in-2019-people-planet-prosperity-in-focus/. (Accessed 19 March 2024), 2021.
[55] K. Brown, K. Rasmussen, The sustainable development goals in 2019: people, planet, prosperity in focus, https://unfoundation.org/blog/post/the-sustainable-development-goals-in-2019-people-planet-prosperity-in-focus/. (Accessed 19 March 2024), 2019.
[56] UN, Do you know all 17 sdgs?, https://sdgs.un.org/goals. (Accessed 19 March 2024), 2015.
[57] M.F. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137, https://doi.org/10.1108/eb046814.
[58] Á. Szaller, G. Pedone, P. Egri Szalóki, G. Nick, A mutualistic framework for sustainable capacity sharing in manufacturing, in: 53rd CIRP Conference on Manufacturing Systems 2020, Proc. CIRP 93 (2020) 938–943, https://doi.org/10.1016/j.procir.2020.04.024.
[59] T.A.M. Tolio, L. Monostori, J. Váncza, O. Sauer, Platform-based manufacturing, CIRP Ann. 72 (2) (2023) 697–723, https://doi.org/10.1016/j.cirp.2023.04.091.
[60] N.R. Magliocca, Intersecting security, equity, and sustainability for transformation in the Anthropocene, Anthropocene (2023), https://doi.org/10.1016/j.ancene.2023.100396 100396.