



Docs4Design Thesis Text Analysis (O3/4)

Wilhelm F. van der Vegte¹

¹ *TU Delft*

This document falls under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Erasmus+ project DoCS4Design (Doctoral Courses System for Design)

The dataset of PhD theses has been published as open data on Zenodo: www.doi.org/10.5281/zenodo.10548837

Citation:

van der Vegte, W.F. (2023). Docs4Design Thesis Text Analysis. DoCS4Design project. Available at <http://DoCS4Design.eu>

12 Dec 2023



DoCS4Design



Doctoral Courses System for Design

Introduction.....	3
Method.....	3
Text preprocessing.....	5
Preparation for clustering by means of quantification and dimension reduction.....	7
Clustering.....	8
Validation of clustering outcomes.....	13
Further analysis of clustering results.....	19
Discussion.....	24
Conclusions.....	26
Credits.....	28
Appendices.....	29
Appendix 1: Clustering results from the different approaches with different numbers of clusters, reduced to two dimensions.....	29
Appendix 2. Orange workflows.....	38









Introduction

Two different approaches were taken to cluster PhD-research content. Firstly, the collection of most-mentioned keywords was taxonomized based on assessment by human experts. This reveals relations between keywords, but not necessarily between PhD projects or dissertations: a dissertation may have some keywords from one coherent group in common with a particular other dissertation, but share related keywords from another group with yet another dissertation. Another way of clustering is applying text-mining algorithms and unsupervised machine learning to group dissertations based on similarities in terms of words used, and from that derive (key)words characterising each group/cluster. PhD candidates who, based on most characteristic words, identify themselves with the same cluster based on most characteristic words might benefit from getting in touch, and look for collaboration opportunities as mates of buddies.

Method

We started out from the database with metadata about 352 theses as presented in the first-iteration report by Stappers, Figoli and Mattioli (Tab. 1).

Tab. 1: Number of 352 theses collected from the six schools

Number	Color	Institution
48		Aalto University
7		Carnegie Mellon University
6		Illinois Institute of Technology
20		Imperial College London
129		Politecnico di Milano
142		TU Delft
352		All institutions



DoCS4Design



Doctoral Courses System for Design

The approach included the following steps:

- Text preprocessing
- Preparation for clustering, using quantification and dimension reduction by applying various approaches
- Clustering by applying different approaches
- Evaluation and internal validation of combined preparation-clustering approaches
- Preparation for external validation
- Further analysis of clustering results.

To implement the workflow defined by these steps, we used the open-source data mining package Orange¹, that is developed and maintained by the University of Ljubljana.

¹ <https://orangedatamining.com/>

Text preprocessing

In our dataset, first we combined the title, the keywords and the entire abstracts for each dissertation and incorporated them into a so-called corpus, where composite keywords ('design for emotion') could be kept together, however only in the part that was copied from the keywords. The corpus comprises one or more columns in a table (or matrix or spreadsheet) where each record (dissertation) has its own row. Apart from the corpus column (one in our case, "title + keywords + abstract" merged into one text), the table has columns for metadata: a unique identifier (ID) for each record, the name of the author, the university (institution) and the year of publishing.

We filtered the corpus to remove common words that are to be considered insignificant and should be ignored when looking for similarities between the records. To that end, we started out from various lists of so-called stop words that can be found on the internet. However, these general lists comprehensively cover very common words like 'the', 'we' 'and', etc., but not words common for a specific type of documents, such as, in our case dissertations in the field of design research, and are not useful for finding clusters or other patterns in the corpus. A common approach to find additional stop words is to visualize the results of preprocessing in a word cloud (Fig. 1), checking it for potentially non-discriminative words, and iteratively expand the list of stop words, while keeping track of which words we did not want to include in our analysis. For instance, a typical formulation often found in the abstract of a dissertation is something along the line of "This thesis/dissertation is structured as follows: Chapter 1 introduces (...), Chapter 2 presents (...)". If we would not filter out words like thesis, dissertation and chapter, dissertations who use this formulation might end up in the same cluster just because of using these words. The same applies to words like literature, abstract, research and questions

Other typical words that were added to the list of stop words are:

- Words typically occurring in dissertations, such as thesis, dissertation and chapter;
- Names of universities: theses from the same institution should not end up in the same cluster just because of that. On the contrary, it is more meaningful to find similarities in projects across institutions.



DoCS4Design



Doctoral Courses System for Design

- Geographical indications such as London, Finland, Europe, because they are likely to have a correlation with one or more specific institutions.
- Names of companies, for the same reason.

We decided not to filter out the word 'design' since a substantial number of theses did not mention the word 'design' in their corpus text, and it could be meaningful that they have this in common. As a final preparation step, we filtered by part of speech and continued with nouns, verbs, and adjectives only.



Fig. 1. Final word cloud after iterative preprocessing. Note that 'understanding' and 'understand' appear together because 'understanding' refers to the noun, while 'understand' refers to all forms of the verb, including 'understanding' as a gerund)

Preparation for clustering by means of quantification and dimension reduction

The first step after preprocessing is to capture the corpus in numerical features (independent variables) based on which algorithms can find similarities. The most common way to do this is to model the corpus as a so-called bag of words (BoW). Put simply, BoW adds a column to the data table for each word in the entire corpus, with for each record a number that expresses its importance in the document, usually based on inverse document frequency (IDF). The number is zero if the word does not appear in the document and otherwise a number between zero and one. The result is a matrix with many zeroes, or sparse matrix, which is typically represented in a compressed notation. Two alternatives that avoid the necessity of a sparse-matrix notation are document embedding (DE) and similarity hashing (SH), which convert each document to an abstract multidimensional vector, with each component a column in the data table.

Three common ways to quantitatively map and visualize the similarities based on the extension of the data table produced by BoW, DE and SH are a distance matrix (DM), and two techniques for dimension reduction: multidimensional scaling (MDS) and t-distributed stochastic neighbour embedding (t-SNE), which is in fact a more recently developed alternative to MDS. MDS can also be performed on the output of a distance matrix.

Clustering

For the actual clustering, we have focused on two options:

- hierarchical clustering (HC), which requires a distance matrix as input (based on cosine distances, as recommended for text), and
- Gaussian mixture models visualized in an annotated corpus Map (ACM), which is tailored for use after dimension reduction with MDS or t-SNE and based on the LabelTransfer algorithm².

Two common clustering algorithms, k-Means clustering and DBSCAN have also been considered, but these turned out not to be able to distinguish more than two clusters, or sometimes even only one. The same is also true in the dashed GMM clustering options shown in Fig. 2, which presents an overview of the considered workflows. The dashed HC clustering options were also disregarded since they produced extremely unevenly distributed numbers of clusters, e.g., one large cluster comprising more than half of all the dissertations together with other clusters of only one or two dissertations.

Ultimately, four approaches were implemented to arrive at different clusterings:

- Hierarchical clustering based on distances from Bag of Words;
- Gaussian Mixture Models based on MDS applied to distances from Bag of Words;
- Gaussian Mixture Models based on t-SNE applied to Bag-of-Words features;
- Gaussian Mixture Models based on t-SNE applied to document embedding.

² <https://orangedatamining.com/widget-catalog/text-mining/annotator/>

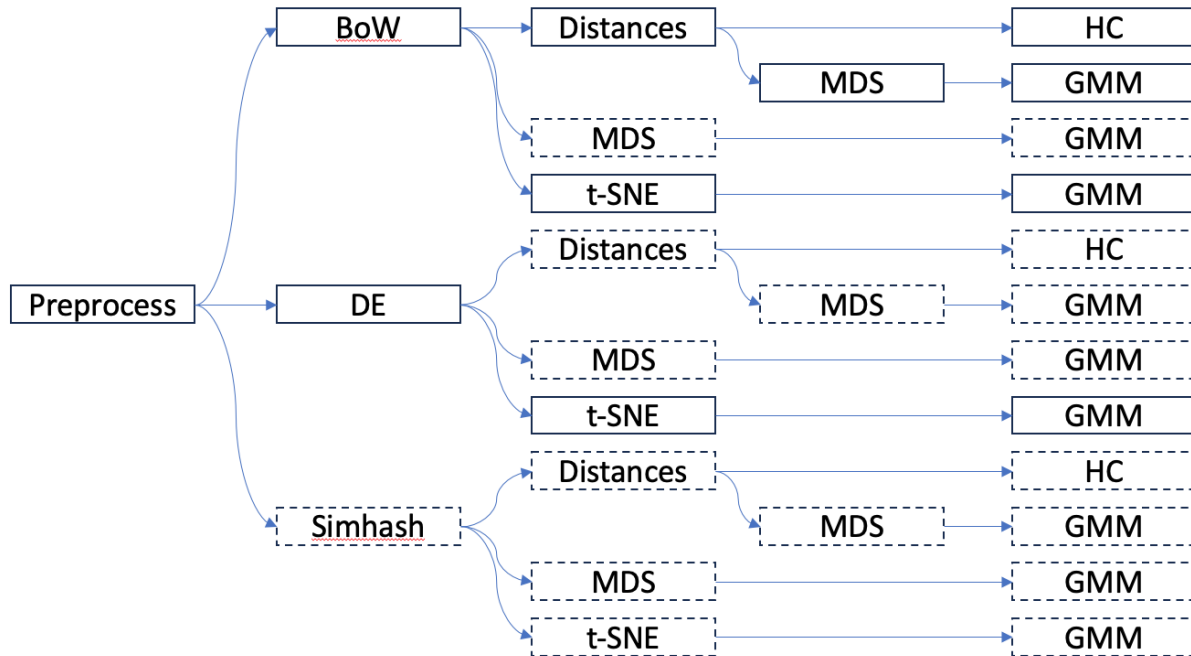


Fig. 2. Possible workflows towards clustering. Dashed options were disregarded due to unusable outcomes.

As an example, Fig. 3 partially visualises the result of hierarchical clustering resulting in 14 clusters. The number of clusters can interactively be increased by dragging the dashed vertical line to the right, and decreased by moving it to the left.

As an example of output of the workflows ending with GMM, Fig. 4 shows the result of BoW → t-SNE → GMM with 9 clusters. The other GMM results also show clusters with hulls (dashed envelopes) and five characteristic (key)words per cluster.

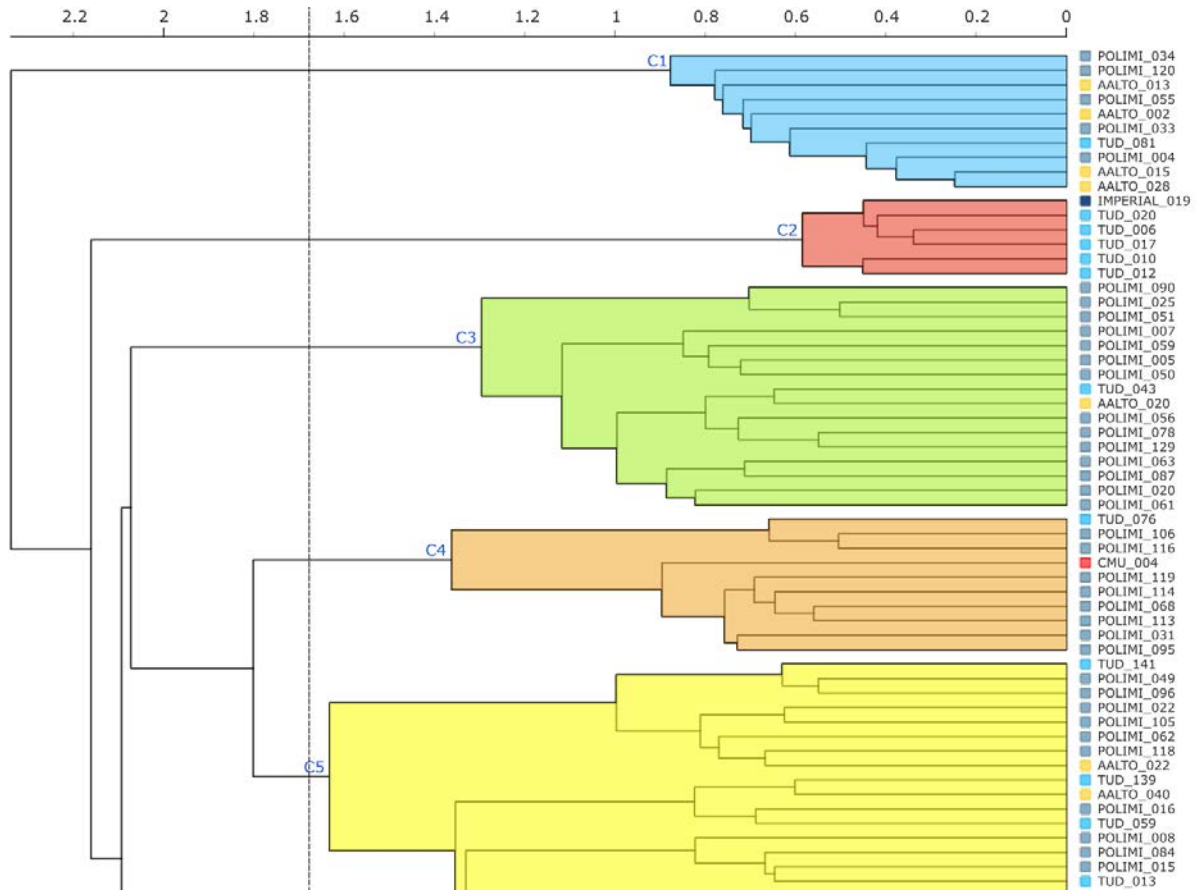


Fig. 3. Partial visualisation of hierarchical clustering output.



DoCS4Design

Doctoral Courses System for Design

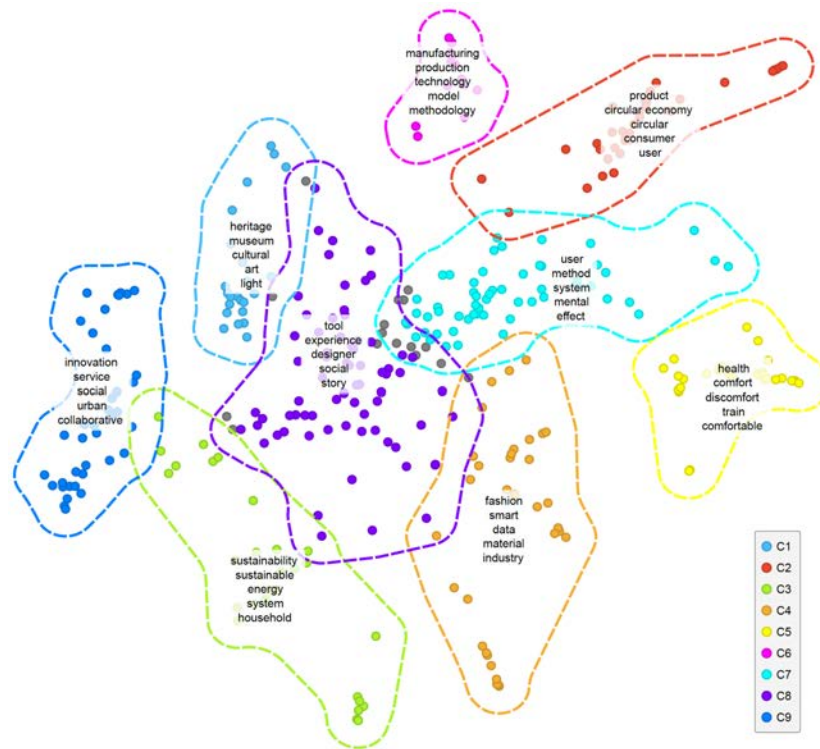


Fig. 4. Example of BoW → t-SNE → GMM output, in this case set to produce 9 clusters. See Appendix 1 for all results.

HC and GMM have in common that they allow the user to prescribe the total number of clusters to be distinguished in the dataset. Since PhD candidates should have enough diversity in clusters to choose from, and on the other hand not be overwhelmed by the number of different options, we aimed to present between 8 and 20 clusters. As already mentioned, in some cases HC produces clustering results with very unevenly sized clusters (i.e., number of dissertations per cluster). This is expressed by the standard deviation in the cluster sizes. When stepwise increasing the number of hierarchical clusters from $N_{clusters} = 8$ to $N_{clusters} = 20$, the standard deviation varies, showing local minima and maxima.

We used this variation in quality of clustering to select 'optimal' numbers of clusters for each approach. We calculated the standard deviation relative to the average cluster size, since a standard deviation of, say, $\sigma = 2$, implies more variation if the average $n_{average} = 3$ than if $n_{average} = 10$. Thus, $\sigma_{relative} = \sigma / n_{average}$. For the HC clustering workflow, local minima of $\sigma_{relative}$

occur at $N_{clusters} = 14$ and $N_{clusters} = 20$, so we selected these clustering results for further investigation.

For the GMM clustering results, we also considered the numbers of dissertations that could not be clustered. To combine these with the standard deviations in cluster size, both numbers were normalized to the interval [0,1] across the clustering results ranging from 8 to 20 clusters and averaged, resulting in a score between 0 and 1 expressing both aspects of clustering quality. Tab. 2 and Fig. 5 show this for one of the GMM clustering workflows.

Tab. 2. Selection of 'optimal' numbers of clusters for the clustering workflow DE - t-SNE - GMM, with local minima at 13 and 17 clusters (highlighted), considering variability in cluster size and number of dissertations that could not be clustered.

$N_{clusters}$	$n_{unclustered}$	$n_{average}$ per cluster	σ	$\sigma_{relative}$	$n_{unclustered}$ normalized	$\sigma_{relative}$ normalized	average of normalized values
8	27	40.63	18.80	0.46	0.84	0.81	0.83
9	32	35.56	15.87	0.45	1.00	0.78	0.89
10	31	32.10	7.95	0.25	0.97	0.43	0.70
11	30	29.27	10.76	0.37	0.94	0.65	0.79
12	29	26.92	8.47	0.31	0.91	0.55	0.73
13	15	25.92	9.13	0.35	0.47	0.62	0.54
14	18	23.86	8.02	0.34	0.56	0.59	0.58
15	23	21.93	9.26	0.42	0.72	0.74	0.73
16	16	21.00	10.07	0.48	0.50	0.84	0.67
17	13	19.94	10.48	0.53	0.41	0.92	0.66
18	21	18.39	10.05	0.55	0.66	0.96	0.81
19	21	17.42	8.88	0.51	0.66	0.90	0.78
20	22	16.50	9.40	0.57	0.69	1.00	0.84

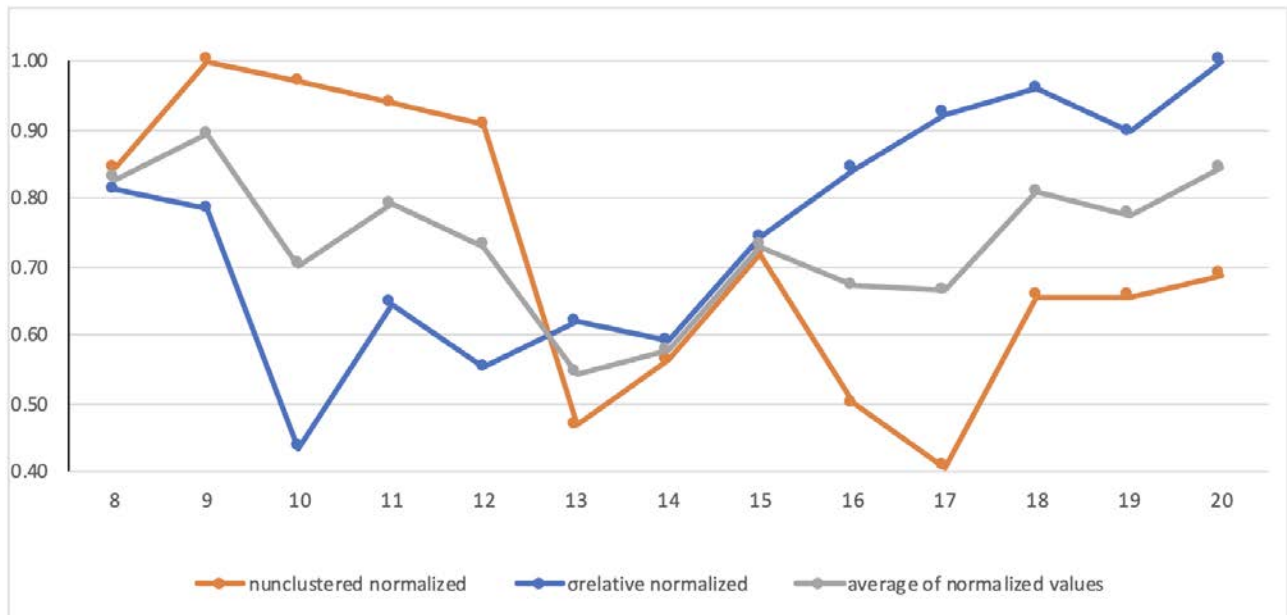


Fig. 5. Graphical representation of the last three columns in Tab. 1.

Based on this approach, we ended up with eight clustering results, two per workflow with different total numbers of clusters: BoW → distances → HC: 14 clusters and 20 clusters;

- BoW → distances → MDS → GMM: 11 clusters and 19 clusters;
- BoW → t-SNE → GMM: 9 clusters and 20 clusters;
- DE → t-SNE → GMM: 13 clusters and 17 clusters.

For each outcome, we extracted five characteristic words per cluster as provided by the Annotated Corpus Map, which were used in the initial external validation presented in the next section.

Validation of clustering outcomes

A common approach to internal validation of clustering results is silhouette analysis. Briefly summarized, for each record in a clustering result a silhouette score can be calculated that expresses, at the same time, how close the record is to other records in the same cluster and how far it is from records in other clusters. The score ranges from -1 to +1, with a positive value

expressing that the record fits well in its own cluster, and a negative value that the value is an outlier. Averaging the silhouette score over all records in each cluster for each of the eight clustering outcomes, then averaging the average score per cluster over all clusters gives a good impression of how successful the clustering effort has been. However, silhouette scores are also known to be less suitable in cases where clusters are irregularly shaped clusters as they often result from GMM.

Therefore, we also applied an initial external validation through a survey in which experts in the field of design research from the 6 institutions were asked to rate each clustering on a five-point Likert scale ranging from 'strongly disagree' to 'strongly agree' according to the statements (1) The "typical words" highlighted for each cluster correspond to coherent sub-areas of design-related research and (2) The visualisation shows closely related sub-areas next to each other and less-related sub-areas further apart. Each clustering was shown only with its five characteristic (key)words and around a centroid (central point of the cluster) in 2D-space, since we did not want the numbers and colours of dots and the cluster hulls to have influence on the scoring. Fig. 6 shows how the clustering in Fig. 4 was presented to the participants. Since hierarchical clustering does not provide a 2-axis plot, we had to implement a workaround by applying MDS to the distances it was based on and using an annotated corpus map to show the 5 most characteristic words in 2D space of the MDS x and y coordinates (See Appendix 1.1 / 1.2). This does not accurately reflect the hierarchical clustering and therefore it was to be expected that the two HC results would not score well on the second question.

In total, 18 experts participated, one of which only scored statement (1) for one clustering. Out of these 18 experts, 7 were from Politecnico di Milano, 4 from TU Delft, 2 from Aalto, 2 from Imperial College, 1 from CMU, and 3 did not specify their employer.

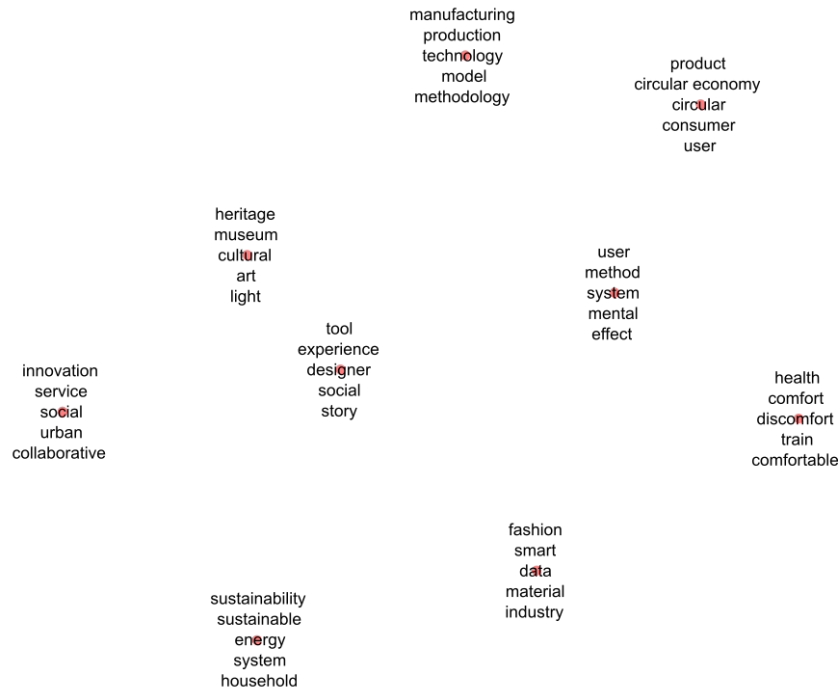


Fig. 6. Clustering from Fig. 4, simplified for ranking.

Tab. 4. Clustering validation results. For all scores, -1 is worst, 0 is neutral and 1 is best. Likert scales in external validation were averaged over the respondents, with 'strongly disagree' mapping to -1 and 'strongly agree' mapping to 1.

Clustering workflow	Internal val.	Initial external validation	
	Average silhouette score	Characteristic words	Distances between clusters
BoW → distances → HC: 14 clusters	0.223	0.25	0.14
BoW → distances → HC: 20 clusters	0.194	0.19	0.14
BoW → distances → MDS → GMM: 11 clusters	-0.168	0.16	-0.12
BoW → distances → MDS → GMM: 19 clusters	-0.156	0.15	0.16
BoW → t-SNE → GMM: 9 clusters	-0.050	0.25	-0.37
BoW → t-SNE → GMM: 20 clusters	-0.110	0.16	-0.22
DE → t-SNE → GMM: 13 clusters	-0.067	0.04	0.00
DE → t-SNE → GMM: 17 clusters	-0.064	0.06	0.24

Tab. 4 shows the results of the internal and the initial external validation. From these, we can conclude the following:

- On average, all clusterings scored positive on characteristic words corresponding to coherent sub-areas of design-related research. However, some scores are unconvincingly close to 0, and BoW → distances → HC into 14 clusters together with BoW → t-SNE → GMM into 9 clusters scores best. The latter however is unconvincing when it comes to mapping the distances between clusters. The fact that HC into 20 clusters scores third-best on characteristic words and similar to 14 hierarchical clusters on distances, seems to indicate that HC gives good results in general.
- Even when we disregard the silhouette scoring as internal validation, which expectedly does not work out well for GMM-based clustering, the other results are a mixed bag, and therefore not the most convincing.

At first sight, BoW → distances → HC into 14 clusters seem the most promising candidate for offering PhD candidates a matchmaking tool. However, at closer inspection, this clustering has one very large cluster, C14, comprising 148 out of 352 dissertations, and another large one, C13, of 80 dissertations, which is still a rather imbalanced distribution despite the initial search for local minima in standard deviation in cluster size. The way hierarchical clustering works is that each time the number of clusters is increased by one, one of the clusters is decomposed into two. Tab. 5 illustrates how the decomposition of clusters changes stepwise from 14 to 20 clusters.



DoCS4Design

Doctoral Courses System for Design

Tab. 5. Hierarchical clustering decomposition from 14 to 20 clusters

#clusters:	14	15	16	17	18	19	20
cluster decomposition	C1	C1	C1	C1	C1	C1	C1
	C2	C2	C2	C2	C2	C2	C2
	C3	C3	C3	C3	C3	C3	C3
	C4	C4	C4	C4	C4	C4	C4
	C5	C5	C5	C5	C5	C5	C5
				C6	C6	C6	C6
	C6	C6	C6	C7	C7	C7	C7
	C7	C7	C7	C8	C8	C8	C8
	C8	C8	C8	C9	C9	C9	C9
	C9	C9	C9	C10	C10	C10	C10
			C10	C11	C11	C11	C11
	C10	C10	C11	C12	C12	C13	C13
	C11	C11	C12	C13	C13	C14	C14
	C12	C12	C13	C14	C14	C15	C15
C13	C13	C14	C15	C15	C16	C16	
C14	C14	C15	C16	C16	C17	C17	
	C15	C16	C17	C17	C18	C18	
				C18	C19	C19	
	C19	C20	C20	C20			

The table shows that when increasing the number of clusters from 14 to 20, it is exactly the large cluster C14 that is affected most by further decomposing into more clusters. Knowing how the 20 clusters are derived from the better-performing subdivision into 14 clusters, we select BoW → distances → HC into 20 clusters as the best candidate for performing a final external validation that has yet to take place, even though we still end up two large clusters C16 with 80 dissertations (the former C13 that stays untouched) and C20 with 84 dissertations. Based on Tab. 5, we can take into account that the groups C5-C6, C10-C12 and C17-C20 might be at risk of creating too much distinction. Fig. 7 shows the selected clustering in the same manner as Fig. 6, with cluster numbering added.



DoCS4Design

Doctoral Courses System for Design

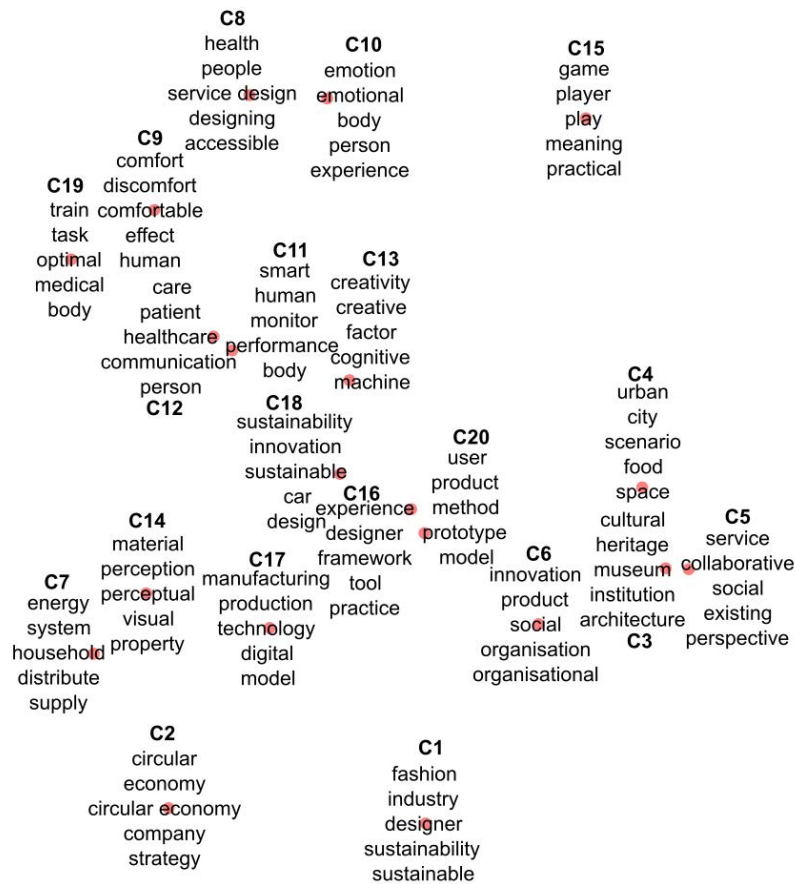


Fig. 7. Selected clustering BoW → distances → HC into 20 clusters as shown in initial external validation, with cluster numbers added (cf. Fig. 6)

An interpretation of what these clusters mean is offered in the next section, together with additional data, such as size of the clusters, average publication year, and presence of each institution in each cluster.

In the final external validation by PhD candidates currently working on their project, we aim to let them select the cluster that they best identify with based on (key)words. Based on that cluster, we will randomly present to them a number of dissertation titles from our database: n titles from the same cluster, n titles from the farthest other cluster and n titles from the nearest other cluster,

and let the participants rank the relevance of each title³. The goal is to test the hypotheses that (1) on average, titles from the 'own' cluster score better than titles from the farthest cluster (weak hypothesis) and (2) on average, titles from the 'own' cluster score better than titles from the nearest cluster (strong hypothesis). In both cases, the null hypothesis is that there is no difference in average scoring of the titles. If at least the weak hypothesis is shown to be true and the null hypothesis rejected, we expect that our definitions of clusters can be deployed to help PhD candidates find 'buddies' in other institutions who work on related topics.

Since the provisional workaround we applied to map the HC results in 2D space turned out not to be optimal, it is probably worthwhile to consider other distance measures to determine the nearest and farthest other cluster for each cluster. Most likely, a good approach is to simply average the already computed cosine distances between each dissertation in cluster X and each dissertation in cluster Y to define the distance between each combination of clusters (X,Y).

Further analysis of clustering results

Apart from the fact that, as suggested in the previous section, the outputs of some approaches lead to clusters that correspond to what is recognised by experts in the field as coherent research areas, purely based on words, additional interesting patterns might emerge by also taking into account other metadata from the database. To reveal these, we collected the following statistics per cluster:

- Number of dissertations in each cluster.
- 20 most relevant words, from most to least frequent. To this end, we applied TDF-IDF⁴-based keyword extraction in Orange, which produces a set of characteristic words that is slightly different from the keywords extracted before, in an annotated corpus map using a threshold based on corrected p-value (false-discovery rate, FDR).
- Average and median publication year and standard deviations of these over all clusters.
- Share of a cluster among all dissertations per institution.

³ In addition, abstracts could be provided as well.

⁴ term frequency weighted by inverse document frequency



DoCS4Design



Doctoral Courses System for Design

- Distribution of dissertations over the six institutions in percent, and standard deviation of the distribution per cluster.

Tab. 6 gives the results in a table sorted from largest to smallest cluster, with heat-mapping applied to the remaining quantities.

Tab. 6. Overview of characteristics of the 20 clusters

Cluster	20 most relevant words	# of dissertations in cluster	Publication year		% of own theses					% of total in cluster					standard deviation over institutions per cluster		
			Average	Median	Polimi	Delft	Aalto	Imperial	CMU	IIT	Polimi	Delft	Aalto	Imperial		CMU	IIT
C20	user product designer method model prototype system experience methodology technology share design mental understanding activity performance home process light development	84	2015.86	2016	16	32	19	50	0	0	24	54	11	12	0	0	20.3
C16	experience practice space tool project narrative data social colour service user designer process digital product personal designing object language theory	80	2016.61	2017	33	11	35	10	29	17	54	19	21	3	3	1	20.2
C18	sustainability design car product future innovation object system ergonomics business approach driving sustainable transition culture development driver interior education market	47	2017.21	2018	6	18	19	0	29	33	17	55	19	0	4	4	20.4
C6	social product organisational design service process innovation phase implementation sustainable knowledge systemic model journey transfer organisation sector company utilize public	18	2016.72	2018	5	5	2	10	29	0	33	39	6	11	11	0	15.6
C3	museum collection art model digital institution architecture audience building product participatory process create public experience heritage community content foundation interface	16	2016.44	2016.5	11	1	2	0	0	0	88	6	6	0	0	0	35.1
C19	skill train medical task product physical system device body model social transfer monitor validation knowledge adaptation discomfort control lack team	11	2016.73	2015	1	6	0	5	0	0	9	82	0	9	0	0	32.3
C1	sustainable sustainability technology detail industrial role practice production designer stage expand theory sector life experience explore implement traditional experiment engage	10	2017.7	2018.5	4	1	8	0	0	0	50	10	40	0	0	0	22.5
C4	food scale common collaborative process base light people social explain network innovative strategy participation data consumption pattern space population service	10	2015.5	2014	6	1	0	0	14	0	80	10	0	0	10	0	31.4
C14	perception critical representation project product interaction mode designer tool feature process image light manufacturing method sensory phenomenon practice perceptual visual	9	2018.78	2019	2	4	2	0	0	0	33	56	11	0	0	0	23.1
C5	evaluation interaction interface infrastructure capability student co-design food improvement university e.g workshop initiative innovation intervention mediate mechanism interpersonal local model	8	2014.88	2015	5	1	2	0	0	0	75	12	12	0	0	0	29.3
C12	patient product communication lab emotion living healthcare consumer service environment people field smart explore technology approach learning analysis engineering flow	8	2018.5	2018.5	2	3	2	0	0	0	38	50	12	0	0	0	22.0
C7	smart exchange technology practice sustainable home force role artefact system transition change time household access solution power develop measure habit	7	2017	2017	1	3	2	5	0	0	14	57	14	14	0	0	21.0
C2	process strategy consumer model company perspective sustainable environmental interaction design methodology waste material framework relationship opportunity user set knowledge prototype	6	2020	2020.5	0	4	0	5	0	0	0	83	0	17	0	0	33.2
C8	designing service design improve design patient people process engagement enhance solution system environment performance accessible context building improvement knowledge model result	6	2016.17	2016	0	3	4	0	0	0	0	67	33	0	0	0	28.0
C9	environment property body recommendation interaction built factor design context hand perceive test tool control interior level component sensory effect company	6	2017.67	2017.5	0	4	0	0	0	0	0	100	0	0	0	0	40.8
C10	designer slow positive choice experience validate discuss movement body rich negative technology understanding concern term explore user human image stimulus	6	2015.83	2016.5	2	2	0	0	0	0	50	50	0	0	0	0	25.8
C17	methodology industry surface process skill challenge perception shape support sector analysis individual personal model industrial goal influence interpretation artefact investigate	6	2017.17	2017	3	1	0	0	0	0	67	33	0	0	0	0	28.0
C11	function design generation designer artefact context data performance behaviour social consumer interaction monitor people collective body changing dynamic inform processing	5	2017	2017	1	2	0	0	0	17	20	60	0	0	0	20	23.4
C13	product idea digital architecture image creative influence transition human experimental data support advance capability pattern system taking outcome cognitive produce	5	2017.8	2019	1	0	0	15	0	17	20	0	0	60	0	20	23.4
C15	topic social emotional co-design collaboration theme issue play service player attitude negative creative empirical project category goal integration concept practical	4	2015.5	2015	2	0	2	0	0	17	50	0	25	0	0	25	20.4
			standard deviation	1.24	1.63	8	8	9	11	11	10						

The table reveals a few interesting characteristics of the clusters. Subjectively, the (key)words extracted as shown in Fig. 7 seem to make more sense than the ones extracted based on TDF-IDF. These are also the ones that were externally validated. Therefore, we will primarily use the (key)words from Fig. 7 to refer to clusters content-wise, while also looking at the thesis titles. In the bullet points below, we will start with an attempt to characterize the clusters, from largest to smallest.

- C20 ('user - product - method - prototype -model') is the largest cluster and, intuitively, the most general one. Yet, the American institutions are absent in this cluster. The impression that the topics in the cluster are rather general is supported by the fact that the average and median publication year are close to the middle of the investigated period. Although

Delft has the largest number of theses in this cluster, it seems to be the most prominent cluster among the PhD research done at Imperial.

- C16 is close behind C20 in size. 'Experience' is the key term in this cluster, otherwise it seems also rather general, listing 'designer', 'framework', 'tool' and 'practice'. It is the only cluster in which all institutes are represented, and the share among the research in the institution ranges from 10% to 35%. This cluster is dominated by Polimi, and, in a relative sense, holds a large deal of the theses from Imperial college: 50%.
- C18 is also a larger cluster, spanning 48 theses. It appears to be dominated by two fields that, at least intuitively, hardly seem to be connected. One is related to sustainability and the other is related to cars, transportation and driving. Apart from these two fields, a wide range of other topics is included, such as ergonomics, entrepreneurship and gaming. It is dominated by TU Delft, relatively prominent at IIT ($\frac{1}{3}$ of its theses) and only Imperial College is absent in this cluster.
- If we 'descend' further towards the smaller clusters, their focus seems to become clearer. C6 is already somewhat more focused, with 18 theses on service design, social design and organizational aspects. C3 has a relatively clear focus on museums, cultural heritage and architecture and is dominated by Polimi, both in a relative and in an overall sense. C19 is clearly focused on health-related and medical topics and dominated by TU Delft but also relatively prominent at Imperial College.
- Almost all subsequent clusters that are yet smaller are featuring theses exclusively from 2-3 institutions. C1 has an emphasis on sustainable fashion and is dominated by Polimi and Aalto. C4 is about urbanism and food, connected through urban farming. It is dominated by Polimi but the most prominent at CMU. C14 is mainly about materials, also in relation to perception thereof, and is the most strongly represented in Delft. C5 focuses on social aspects, collaboration and service design, but is otherwise rather unspecific. It is dominated by Polimi. C12 is another healthcare cluster. While C19 seems to focus more on the care-provider or doctor side (e.g., surgery support), C12 seems to focus more on patient experience (e.g., reducing waiting times). TU Delft has the largest number of theses here. C7 addresses the energy transition and is also dominated by Delft. C2 has a clear focus on the circular economy and is strongly dominated by TU Delft. C8 is mostly about

service design in the context of health, again dominated by Delft. C9 has a clear focus on comfort and is populated with theses from Delft only. C10 clearly focuses on emotions and has equal contributions from Polimi and Delft. C17 focuses on manufacturing, mostly additive manufacturing, which seems to be mostly a Polimi topic. C11 revolves around the notion of 'smart' but otherwise it is rather unspecific, addressing smartness areas ranging from cyber-physical systems to mediation of social viscosity. TU Delft has the most contributions here. C13 is about creativity, and mostly digital support thereof. Imperial College, where this seems to be a hot research topic. Finally, C15 shows a clear focus on game and play at Polimi, Aalto and IIT.

- The annotated corpus map in Appendix 1.2 shows the clusters and their hulls after applying MDS to the distances that were used for HC. The figure suggests that the theses in the more focused clusters are also closer to each other in the MDS result, which is evidenced by their much smaller envelopes. In contrast, the large clusters C16, C18 and C20 are practically 'all over the place'.

Some further observations:

- As can be expected, the largest institutions in terms of design-related research, Polimi and TU Delft, contribute to more clusters than the smaller ones. Yet, each institution seems absent in two or more areas: Polimi in C8 and C9 which could be described as medical service design and human comfort, and TU Delft in C13 and C15, which relate to creativity and game play.
- Polimi and TU Delft are also very dominant in specific clusters: Polimi in C3 and C4, which relate to museums and cultural heritage and to urbanism and food; Delft dominates in C2, C9 (100%) and C19, relating to circular-economy strategy, human comfort and medical care.
- Looking at the years of publication, some clusters stand out because they seem to have been popular in the beginning of the investigated period or, on the other hand, are becoming popular towards the end of the period. The clusters C4, C5 and C15 seem to be past their prime: urbanism and food, collaborative service design and game play. The

research in clusters C2, C12, C13 and C14 appears to be relatively new: circular-economy, design for healthcare, creativity & cognition and material perception.

Discussion

If a supposedly objective machine-learning approach is applied to reveal patterns regarding textual content in the PhD dissertations, it seems inappropriate to criticize the results based on human interpretation, as was done in the above attempts to characterize the contents of each cluster. After all, the algorithm may have found similarities that a human would not (easily) extract from the corpus. Yet, the objective fact stands that the clustering effort produced two very large clusters C20 and C16 that are about ten times the size of the majority of other clusters, and one rather large cluster C18 that is about five times that size.

Fig. 8 shows the uneven distribution of cluster sizes. It is not surprising that, content-wise, these larger clusters are perceived as rather unspecific. The initial external validation also produced a score corresponding to 'neutral-to-agree' rather than 'agree-to-strongly-agree' on the Likert scale indicating to what extent the "typical words" highlighted for each cluster correspond to coherent sub-areas of design-related research. If we would have asked our experts to evaluate each individual cluster, they might have revealed that only the smaller clusters were indeed perceived as coherent research areas.

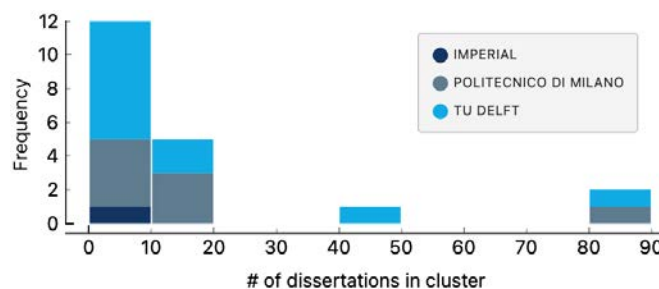


Fig. 8. Distribution of cluster sizes. Frequency corresponds to the number of clusters in each of the size bins. Colours indicate the dominant institution per cluster

We could still test whether the smaller clusters are perceived as more coherent by introducing an additional hypothesis to the final external validation, expressing the expectation that differences in appraisal of the 'own' cluster and other clusters are stronger if the subject chooses a smaller

cluster to identify themselves with. For next dissertation-clustering efforts, confirmation of that hypothesis could suggest that indeed there is a large group of PhD research subjects that is inherently muddled and therefore cannot be assigned to any meaningful cluster, but it could also motivate going further in avoiding large clusters than we already did in the current investigation. Rejection of the additional hypothesis would confirm underlying similarities that are not obvious to the casual human observer.

Three other issues beside unevenly distributed cluster sizes are (1) the sensitivity of clustering results to the choice of stop words to be excluded when looking for similarities, (2) the insensitivity to synonyms and use of the same word in a completely different context and (3) seemingly random misclusterings.

The sensitivity to the selection of stop words emerged when, at an early stage of the clustering efforts, we noticed that expressions like “designer’s” and “user’s” appeared next to “designer” and “user” in the lists of words typical for certain clusters. Since we aimed not to distinguish between the different declensions of words, we performed find-and-replace on the original data to get rid of possessive “s”es. As this initially appeared not to remove all uses of the possessive “s”, we found out that various different apostrophe-like symbols were used to precede the “s”, and that these all had to be copy-pasted from the original data into the list of stop words. This was an iterative exercise where remaining possessive “s”es had to be removed in each next iteration. While doing this, we noticed that between iterations where just another appearance of the possessive “s” was removed, clustering algorithms sometimes suddenly produced very different clusters. This effect may also pertain to our decision not to disregard the word “design”. If we would have disregarded this ubiquitous word, the results would have been very different. In fact, keeping “design” might have been one of the causes of our difficulties getting rid of very large clusters - which in part may have been caused by the fact that all the theses in these clusters mentioned “design”.

In some cases we noticed that some typical words had been used with different meanings that were not recognised as such by the algorithms. This is probably how a dissertation about product care ended up in a health-related cluster, although the care was about product maintenance rather than healthcare. Such misinterpretations may also have caused seemingly random misclusterings where, for instance, two robotics-related dissertations were assigned to clusters to which their membership did not seem obvious, related to (again) healthcare, and the energy

transition, respectively. Such misclassifications may have also contributed to the forming of undesirably large clusters.

An observation that could be made across the different clustering results, is that there were some combinations of at least 3 words that emerged in one cluster as a result multiple approaches:

- 6 out of 8 clustering results held a cluster characterized by the words cultural, heritage and museum, 3 of these in combination with architecture
- The 5 (key)words circular, economy, circular economy, company and strategy appeared together exactly in that same order in 3 out of 8 clusters. This is in part caused by the limitation that the so-called digram “circular economy” could be treated as belonging together when in the list of keywords but not in the title or the abstract, where tokenization had to be done based on spaces. Yet it is interesting to see that exactly the same combination emerged in the two HC-based clustering results as well as in the BoW → t-SNE → GMM result with 20 clusters.

Conclusions

Clustering algorithms can be used to find similarities between dissertations in the field of design, and bring them together into groups sharing similarities. However, it is not always clear which similarities have led to specific groupings. We compared different clustering approaches, and, in our initial external validation, the clustering result that received the highest score on creating meaningful groupings resulted in 20 clusters showing an uneven distribution. Especially for the three largest clusters spanning 40-80 out of the 352 theses it is not obvious why the dissertations in them were treated as similar. This may be due to “hidden” word use patterns that machine-learning algorithms can better distinguish than humans, but perhaps also due to our selection of stop words to be excluded from processing. Most of the 17 smaller clusters spanning 4-18 dissertations each, seem more promising for further investigation. They could more easily be described based on the extracted characteristic words and the contents of the dissertations included in them. This is of course a subjective observation, to be confirmed in a final external validation. Before proceeding to the final external validation, it might also be a good idea to redo the initial external evaluation by letting experts judge the coherence of individual clusters from



DoCS4Design



Doctoral Courses System for Design

different results, and deriving the quality of each clustering approach from the average performance of its constituent clusters instead of an overall judgement per clustering approach.

It has to be noted that text mining is a relatively young field of research, especially when it comes to application to certain specific areas such as dissertations in a particular field. A lot of work has been done on the development, verification and optimization of new algorithms, but there are no best practices as to which clustering approach using what settings should be applied given the size of the corpus (number of theses and number of words in title + keywords + abstract, in our case). What makes finding the best approach even more of a challenge is the fact that, still, new approaches become available that claim to be more sophisticated, more accurate and/or faster. The Annotated Corpus Map that we used in some workflows was actually developed in 2022. We hope, nevertheless, that our current efforts and findings can provide inputs to future work in which, ultimately, best practices might crystallize.



DoCS4Design



Doctoral Courses System for Design

Credits

The analysis of the D4D thesis texts was conducted by TU Delft.

The report was edited by: Wilfred F. van der Vegte.

All partners contributed to the output:

Aalto University

Elise Hodson, Guy Julier, Michel Nader,
Sampsa Hyysalo

Carnegie Mellon University

Jonathan Chapman

Illinois Institute of Technology

Carlos Teixeira

Imperial College London

Rafael Calvo, Weston Baxter

Politecnico di Milano

Annalinda De Rosa, Fabio Figoli, Francesca Mattioli,
Lucia Rampino, Paola Bertola

TU Delft

Annemiek van Boeijen, Erika Hajdu, Pieter Jan
Stappers, Wilfred F. van der Vegte



Appendices

Appendix I: Clustering results from the different approaches with different numbers of clusters, reduced to two dimensions

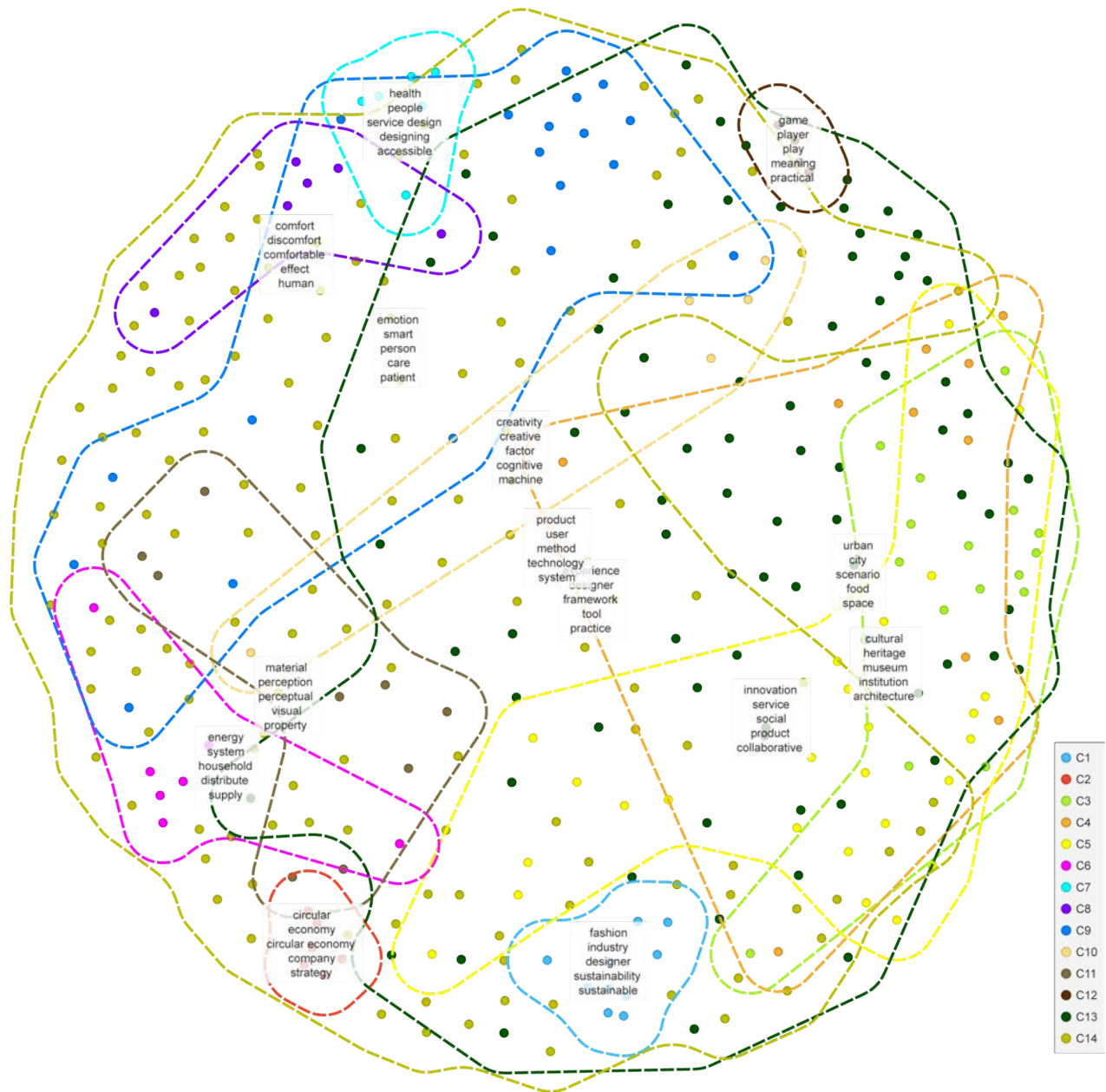


DoCS4Design



Doctoral Courses System for Design

Appendix 1.1 BoW → distances → HC: 14 clusters



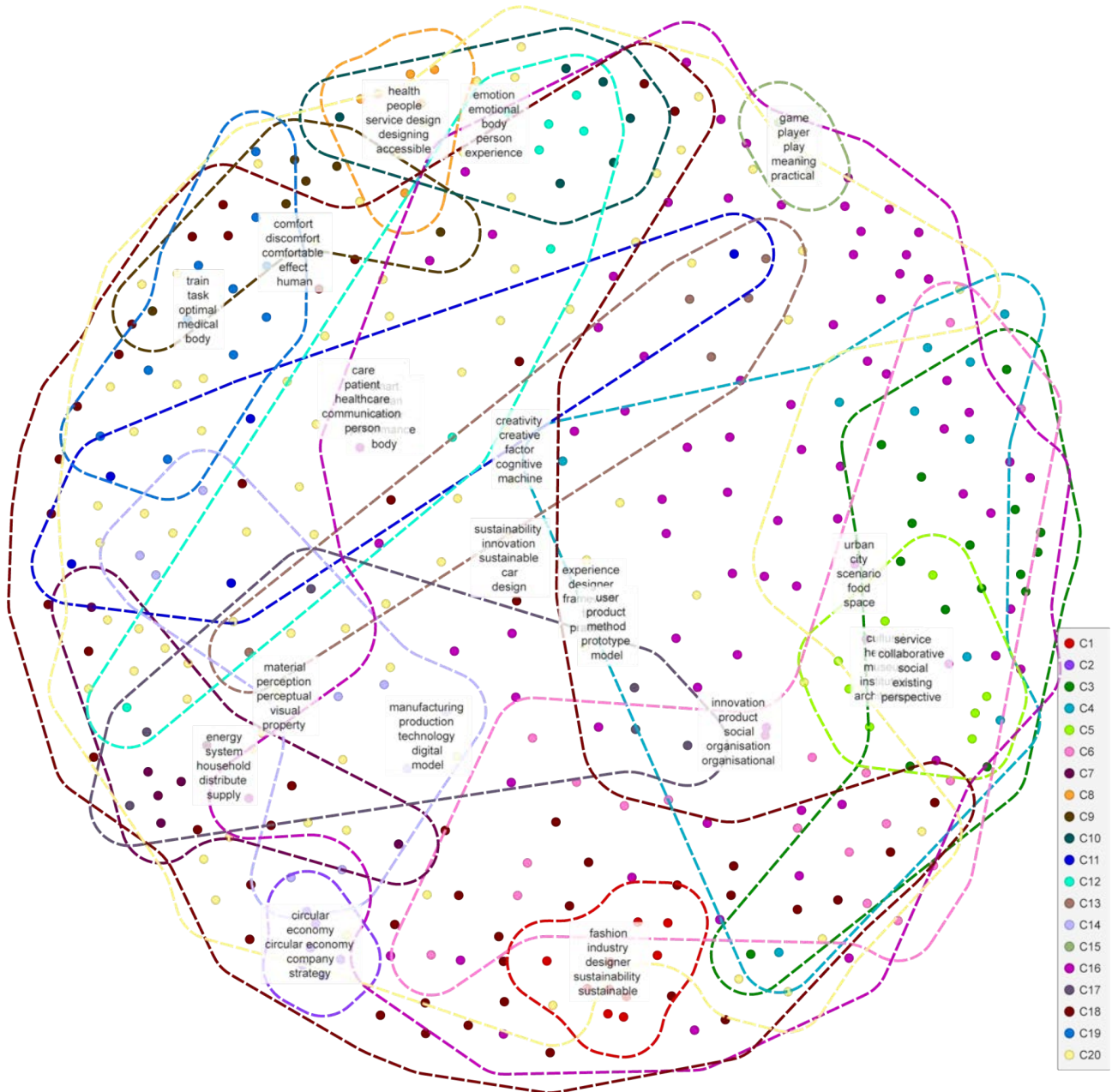


DoCS4Design



Doctoral Courses System for Design

Appendix 1.2 BoW → distances → HC: 20 clusters



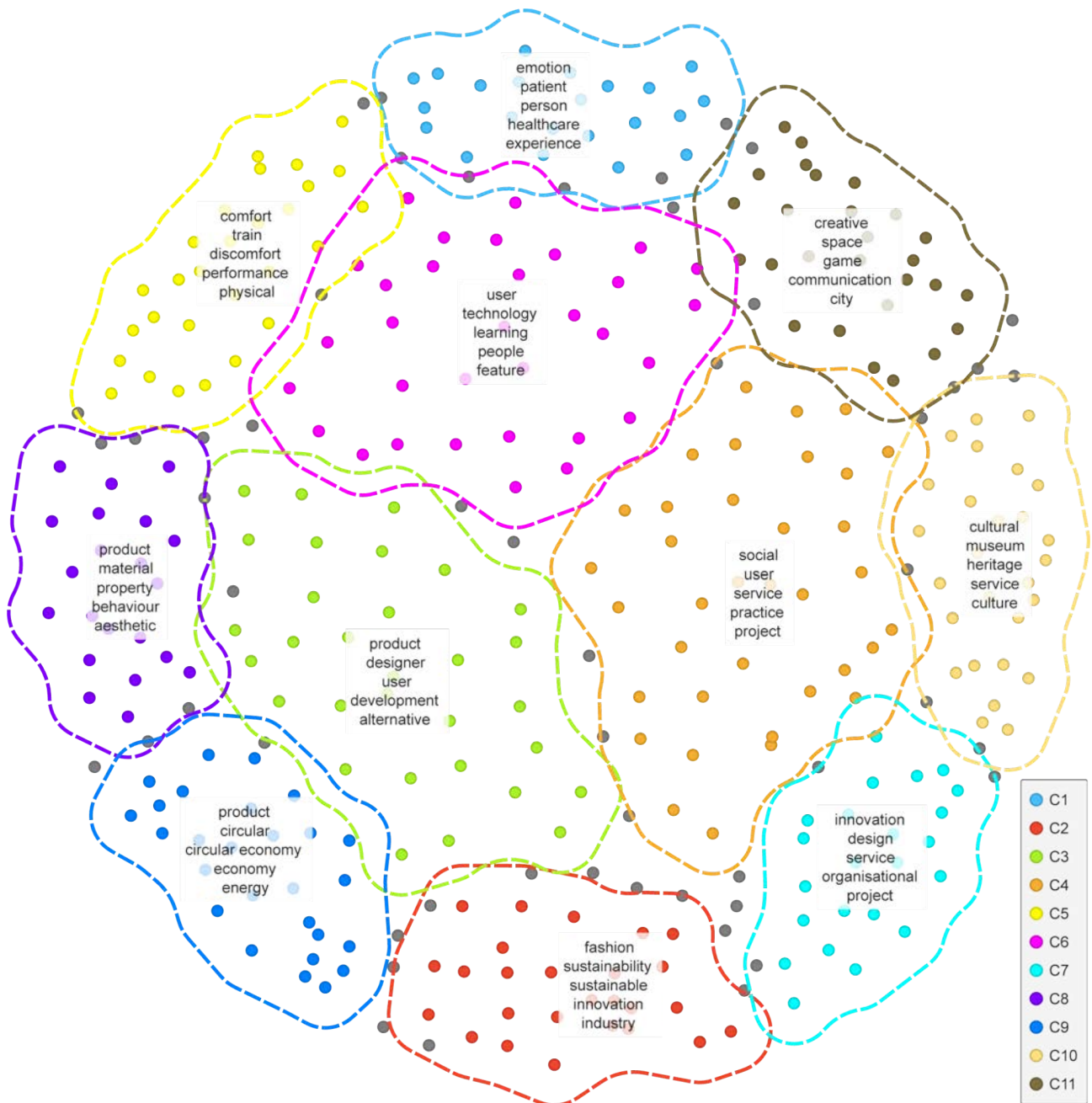


DoCS4Design



Doctoral Courses System for Design

Appendix 1.3 BoW → distances → MDS → GMM: 11 clusters



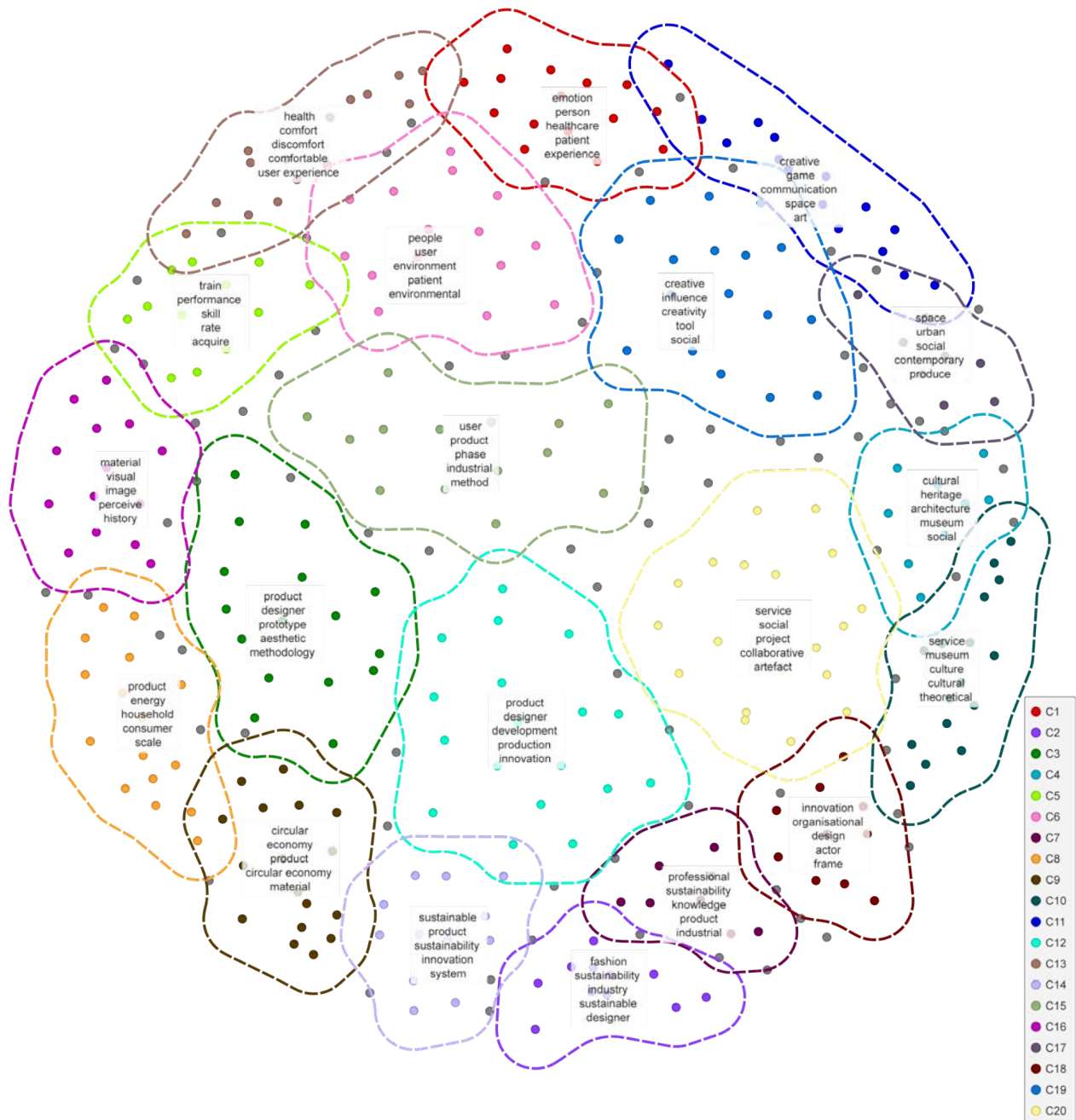


DoCS4Design



Doctoral Courses System for Design

Appendix 1.4 BoW → distances → MDS → GMM: 19 clusters



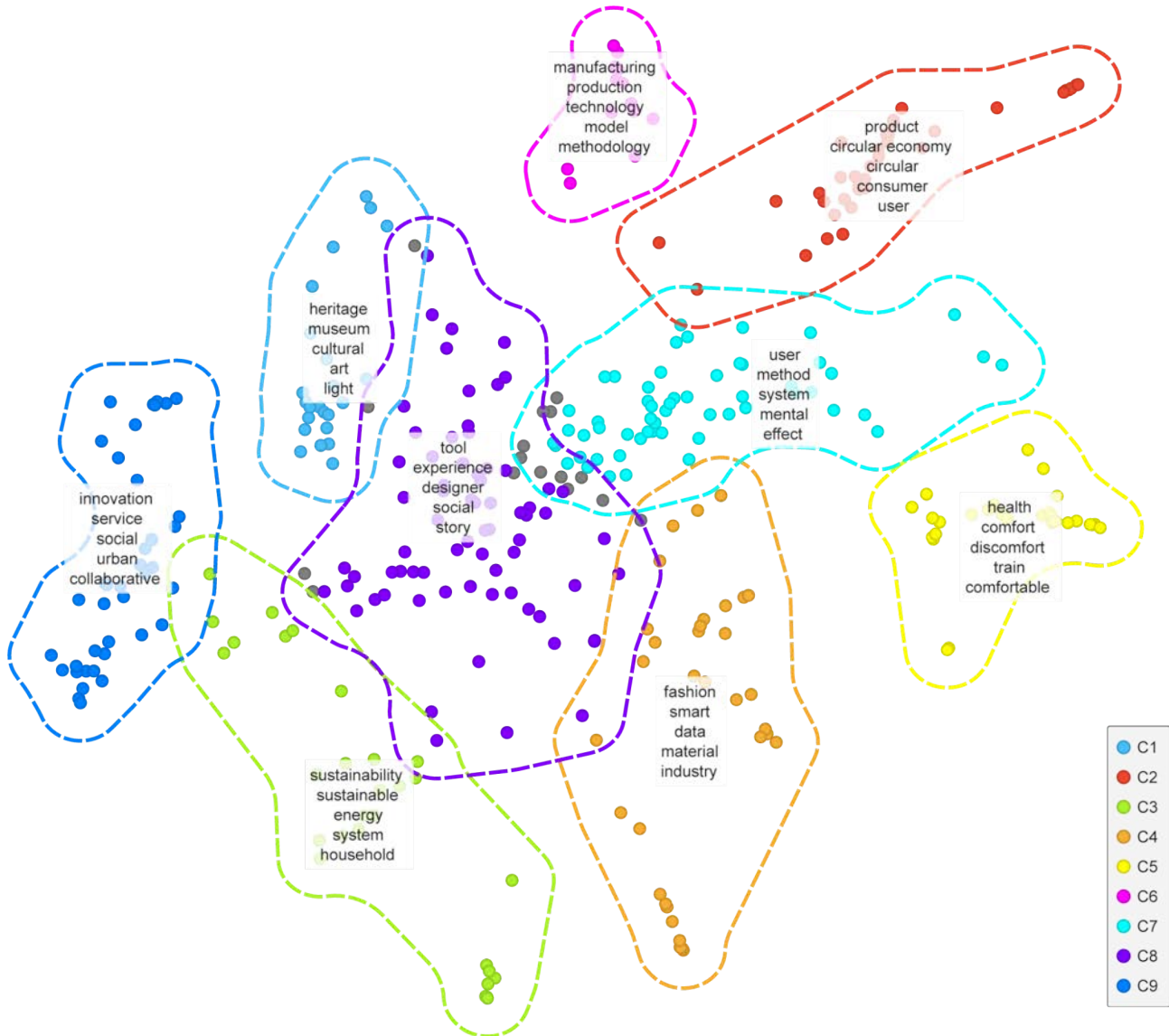


DoCS4Design



Doctoral Courses System for Design

Appendix 1.5 BoW → t-SNE → GMM: 9 clusters





DoCS4Design



Doctoral Courses System for Design

Appendix 1.6 BoW → t-SNE → GMM: 20 clusters



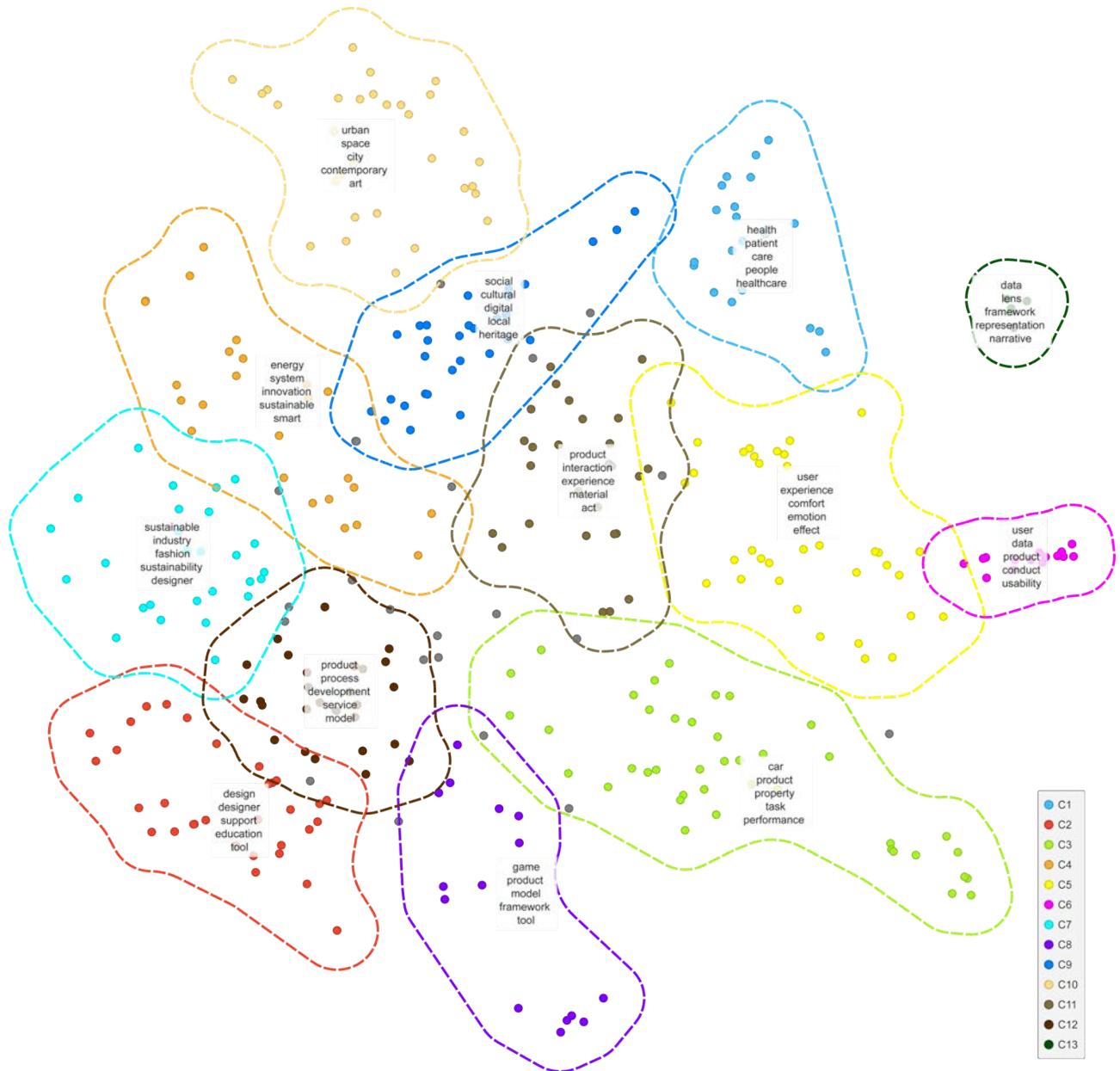


DoCS4Design



Doctoral Courses System for Design

Appendix 1.7 DE → t-SNE → GMM: 13 clusters



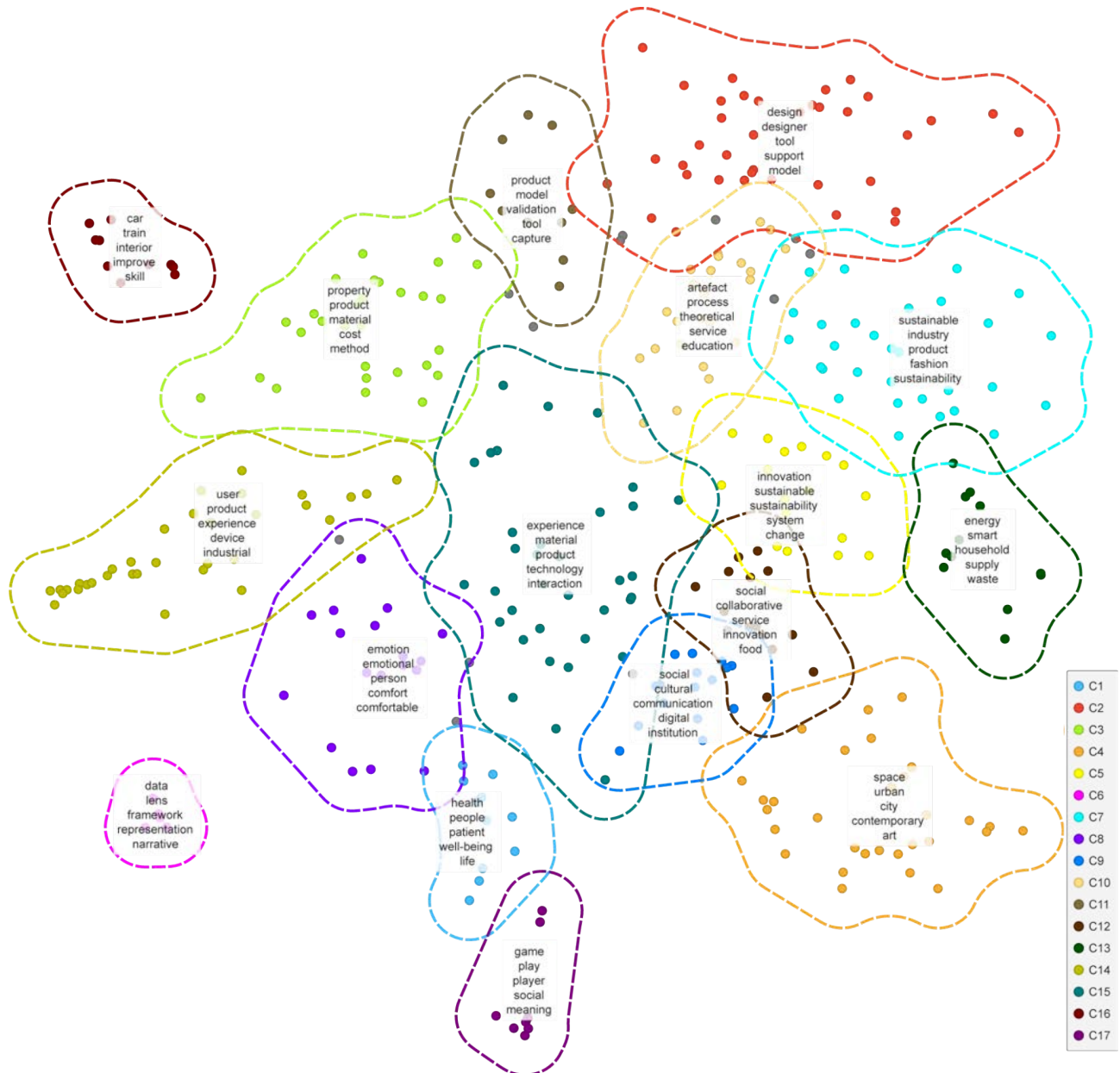


DoCS4Design



Doctoral Courses System for Design

Appendix 1.8 DE → t-SNE → GMM: 17 clusters



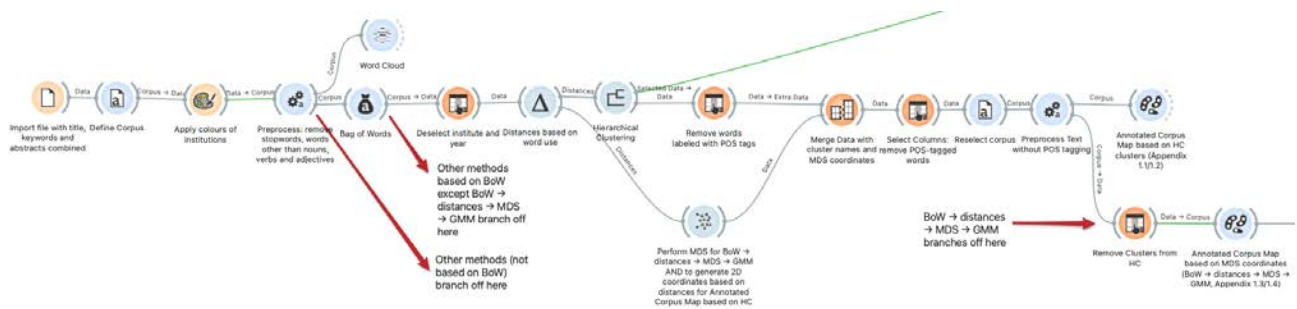


DoCS4Design

Doctoral Courses System for Design

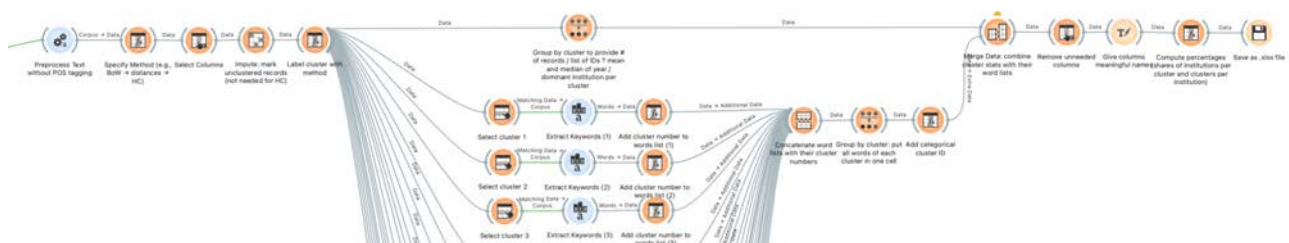
Appendix 2. Orange workflows

Appendix 2.1 Workflow for the methods BoW → distances → HC and BoW → distances → MDS → GMM



This workflow shows also where the other workflows (methods of text clustering) branch off. The lines exiting the workflow connect to further processing to generate Tab. 6 (Appendix 2.2)

Appendix 2.2 Partial workflow to generate Tab. 6



Note: the coloring of the cells based on their values was done using Microsoft Excel.