Enhancing Cross-border Co-operation of Business Organizations based on the Investigation of Textual- and Categorical Information

Ferenc Tolner[1], Balázs Barta[2], Márta Takács[3], and György Eigner[4]

[1]Pannon Business Network Association, Zanati út 32-36, H-9700 Szombathely, Hungary, Óbuda University Doctoral School of Applied Informatics and Applied Mathematics, Budapest, Hungary, ferenc.tolner@am-lab.hu
[2]Pannon Business Network Association, Zanati út 32-36, H-9700 Szombathely, Hungary, balazs.barta@pbn.hu
[3]John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/b, Budapest, H-1034, takacs.marta@nik.uni-obuda.hu
[4]University Research and Innovation Center, Physiological Controls Research Center, Óbuda University, Bécsi út 96/b, Budapest, H-1034, eigner.gyorgy@uni-obuda.hu

Abstract: International cross-border co-operations can highly contribute to individual and even regional value-added generation. In the present work 51 Central European stakeholders with different organizational types were selected for an online survey where preference options in a form of a multiple choice questionnaire were collected on their interests for a possible future co-operation and additional textual data on their general descriptions, strengths, focuses, goals etc. Based on the gathered data clustering of the partakers were performed with the K-modes algorithm for the categorical variables resulted of the questionnaire. Additionally Latent Dirichlet Allocation was used for the textual information in order to present an alternative technique for decision makers to the grouping of business organisations. Such tools can aid more effective and long-term business network formations and contribute to business sustainability and resilience that is of paramount importance in a globalized market framework. This globalized, turbulent environment poses risks that are hardly manageable on an individual level and therefore collaborations and augmentation of business activities gain more and more importance and focus from the side of policy makers and scientific community as well.

Keywords: cross-border co-operation; K-modes; resilience Latent Dirichlet Allocation; topic modeling;

1 Introduction

Due to the current globalization processes just like business organisations, re-search- and educational institutions are also exposed to international competition, challenges and threats caused by the ever growing economic turbulences. Therefore, every possible means of collaboration has to be utilized in order to become resilient against the various business-environmental and organizational fluctuations and remain sustainable in the long-run. Cross-border collaboration and international knowledge sharing may be a proper response to the new challenges, thus the facilitation of opportunities for contact-seeking among actors of various business sectors has to gain momentum in order to ease partnership establishments based on common interests. This latter is the clear goal of regional business cluster formations where ”cluster” in this context means a group of stakeholders that share similar- or even complementary attributes, located geographically close to each other with common goals that are to be achieved by joint activities on specified fields [1].

These industrial- and business clusters enable partakers to extend limits on their fields of specialisation and bring new opportunities closer via information sharing and complementing deficiencies of each cluster members. Being part of a regional- or international cluster can contribute to risk reduction by decreasing the information shortage in business processes and technological related areas, which can easily become a market advantage opposed to other local actors outside of the cluster. Such knowledge sharing – typically present among cluster members – can be especially relevant in case of innovative processes since inter-organization arrangements are pursued on system level and enhanced knowledge generation is expected within the network [2].

Nevertheless, it also has to be noted that information and knowledge is shared among the members of regional clusters better than among actors that are not part of the given cluster. This results in a phenomenon, which characterizes the ”staying in place” of information, which often means the local region where the information was generated. Therefore the enhancement of cross-border partnership establishments beyond regional cluster forming is expressly advised to be promoted from side of policy makers, since resource, knowledge and information sharing is essential for performance increase on an expanded regional level [3], [4], [5].

Business- or industrial clusters are usually formed on regional or national level because of matching individual interests of the founder members or due to outer driving forces or national subsidies. Cross-border, international initiatives however, many times struggle to generate long-term business cluster formations and result in partnership collaborations that are utilised only in short-term projects.

Oftentimes, due to the large geographical distances, different languages, culture- and business customs there is hardly any up-to-date valuable strategic information at hand on remote market opportunities and niche markets. These are on the other hand available or sometimes even obvious at local companies.

Nevertheless, as it is concluded in literature the knowledge generated, shared and disseminated in clusters can get beyond of the individual firms, but generally ”stays” in the regions without benefiting neighbouring or remote territories [3], [4].

The restricted mobility and shortage of strategic information inevitably hinders the economic prospects of individual business actors, local regions and of the European Union as well. Therefore the encouragement of partnership establishments in any form is not just natural, individual business interest than even national and international economic interest. Accordingly, this fact is already addressed by the European Community by supporting innovative cluster forming activities in favour of the overall competitiveness of the regions and also the members of the European Union [2].

In this study our attempt is to investigate methodological options applied on freely or relatively easily obtainable data gathered from Central European innovative business organizations so as to provide more efficient international, cross-border partnership establishments. By achieving this goal, our expectation is that the successful co-operations of organizations of different industrial background will contribute to the overall competitiveness of the Central European region and also increase its resilience against economic turbulences [1].

2 The Collected Data

During our data gathering besides generally accessible address information, organizational type, business branch type and multiple choice preference options were collected in a form of an online survey from 51 innovative Central European stakeholders. The answers were collected regarding optional future co-operation and further education possibilities that the attendees would have been willing to take part in (see Table 1.). The attendees that took part in the questionnaire were from 7 countries and 18 different NUTS2[1] regions from Central Europe (see Subsection 2.1.).

In addition to the multiple choice questionnaire, textual descriptions were also collected that contained besides general information on the organizations (e.g.: fields of current activity) brief descriptions on focuses, strengths etc. Since such data are easy to access from web pages and marketing materials as well, the thorough investigation of their content is supposed to serve with procedures for additional information extraction that can broadly used and generalized (see Subsection 2.1.

[1] Nomenclature of Territorial Units for Statistics: Geocoding standard among the EU member states for referencing the subdivisions of countries for statistical purposes.

Table 1
Possible industrial fields offered in multiple choice selection to map organisation preferences.

Thematic industrial field options for co-operation
Data Analytics, Complex Simulation and Modelling
Machine Vision
Predictive Maintenance
Factory & Process Automation
Digital Innovation & Industries: Machinery
Advanced & Smart Materials
Industrial IoT
Digital Marketing
Innovation in a Circular Economy
Design & Engineering for Additive Manufacturing

2.1 Exploratory Data Analysis

For the exploratory data analysis the Anaconda 3 framework was utilised that is a freely available software environment on Windows, Linux and Mac OS X to simplify scientific computing tasks, the depicted results were generated using the Python 3 programming language [6].

The implemented survey was sent out to organizations that were assumed to be active in innovation and could be interested in knowledge- sharing and generation on international level. Besides industrial- and business organizations Higher Education and Research Institutions, Technology Transfer Institutions, BSOs[2] and DIHs[3] were involved in our survey as well since innovation is present at a higher rate in collaborations where the members are from different organisational types and knowledge sharing is easier when research institutions participate as well [1], [4]. The distribution of organizational types among the partakers in our research is plotted on Fig. 1.

The number of selected options from the multiple choice questionnaire can be seen on the bar plot of Fig. 2. The attendees could mark more than one option for a field of possible future co-operation or further education, therefore a set of categorical variables have resulted whether the given fields were selected or not. The ratio of attendees that selected the given fields are listed on the top of each bar in percentage.

The address information was analyzed with the open source *geopandas 0.8.0* python module in order to gain comprehensive geographical visualisations by geocoding[4] [7]. This enables the geographical investigation of the distribution of attendees regarding their attributes. The geographical location of the investigated 51 organization with the corresponding NUTS2 regions marked can be seen on Fig. 3. From point of view of organizational type it can be concluded that a relatively mixed sampling could be achieved. According to our literature

[2] Business Support Organisations

[3] Digital Innovation Hubs

[4] The process of generating latitude and longitude coordinates and generation of geographic position from addresses and names of locations.

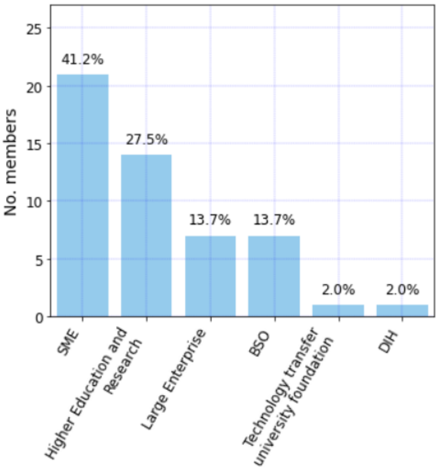


Figure 1
Distribution of organisational types within the investigated sample of 51 organizations.

research, such diffuse geographical distribution might be advantageous for networking efforts. Thus inter-cluster information- and knowledge sharing is supposed to be more intense and proactive in case of bigger distances (less fear of misuse or abuse of information or negative effect on individual market position).

The applicants were also asked for providing data on their industry- or business branch with the corresponding NACE[5] numbers in which they were active at. The members could provide more than one NACE number to cover their field of activities adequately. This general, relatively easy to access information on organizations can give further insight into the sample investigated and could assist to the understanding of the preference choices received from the survey. As can be seen on the pie chart on Fig. 4. the dominant economic activities in the sample given by major NACE categories were *Manufacturing*, *Information and communication* and *Professional, scientific and technical activities*.

The textual information provided by the participants covered the topics listed below (the most common words representing the content of the texts is visualised on Fig. 5. as a *word cloud*):

- General description of attending organisations and major activities.
- An account on main competences and field of expertise. This included their products and special services which constituted the foundations for their economic value-added.
- Short summary on the strengths and outcomes that might entitle the

[5] Standard classification for economic activities used within the European Union.

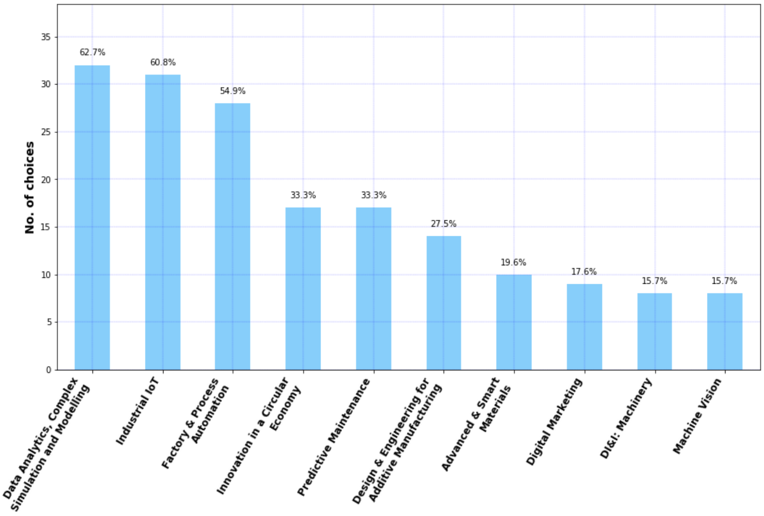


Figure 2
Total numbers and relative ratios of the 51 attendees in the multiple choice selection questionnaire.

partakers to be involved in further co-operation projects.

- Main challenges faced in their corresponding region.
- Brief account on the innovative solutions implemented that contributed to overcome the challenges faced in their corresponding region.
- Key value-added benefit sought by taking part in an international cross-border co-operation activity.

For the topic modeling purposes (elaborated in Subsection 3.1.) a summary box-plot on the number of words in each text has been provided (see Fig. 6.). Since according to literature on topic modeling the length of texts to be analyzed is not all the same for selecting the appropriate topic models [8], [9]. In our case, mostly texts in length of 100 words were available apart from a few exceptions where longer descriptions were given. As texts per participants were concatenated for our topic modeling purposes in the following, there were no text lengths below 50 words, therefore no texts had to be excluded of our textual analysis.

3 Results, Discussion

Partitioning objects into homogeneous groups is a fundamental task of data science. During clustering the separation of the objects are commonly done based on some kind of defined dis(similarity) measure. The expectation towards clustering is to serve with groups in which the objects are more similar to each other with respect to certain characteristics than to other objects ordered to

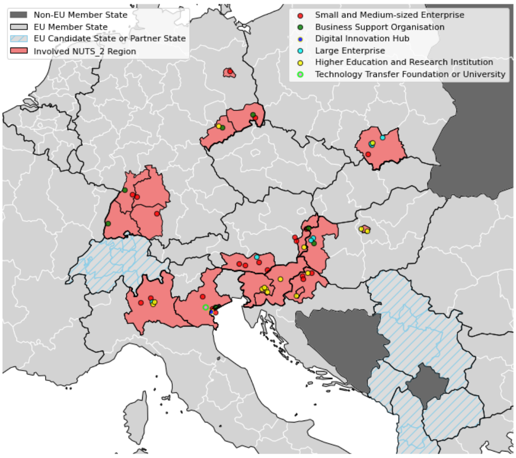

Non-EU Member State
EU Member State
EU Candidate State or Partner State
Involved NUTS_2 Region
Small and Medium-sized Enterprise
Business Support Organisation
Digital Innovation Hub
Large Enterprise
Higher Education and Research Institution
Technology Transfer Foundation or University

Figure 3
The geocoded geographical distribution of the 51 Central European participants marked according to their corresponding organisational types and indication of NUTS2 regions where they are located.

different clusters. This enables the depiction of non-trivial or buried information in an unsupervised way that would be hardly feasible otherwise [10], [11].

3.1 Topic Modeling of Textual Data

Topic modeling belongs to the unsupervised Artificial Intelligence learning methods with broad range of application possibilities. Just like several areas of natural sciences, social studies utilises the options offered by topic modeling techniques as well. Since textual inputs can be obtained more and more easier and often completely free of charge (unlike numeric statistical data on companies like annual revenue, total sales, liquidity ratio etc.) such data has gained an increasing focus in the past decades in characterization of data source subjects. Together with advanced visualisation techniques like *word clouds* or diagrams for topic relations (see Fig. 10.) the comprehension of unstructured large amount of texts has become feasible and obvious in analysing business organizations as well [12].

Latent Dirichlet Allocation (LDA) is one of the most wide-spread and most cited topic modeling techniques that provides a structured overview on the context of large amount of textual data. The method assumes documents to be a discrete probability distribution of latent topics that cannot be observed and likewise the topics to be a discrete probability distribution of words. LDA has been published first in 2003 as a probabilistic model to reveal hidden semantic structures in textual data and since then many authors have generated several results utilizing its potentials in different walks of life [13]. Although LDA has many disadvantages compared to other topic modeling techniques – as it can hardly cope with topic correlations, short text snippets and makes unrealistic assumptions as word- and topic orders does not count–, based on literature

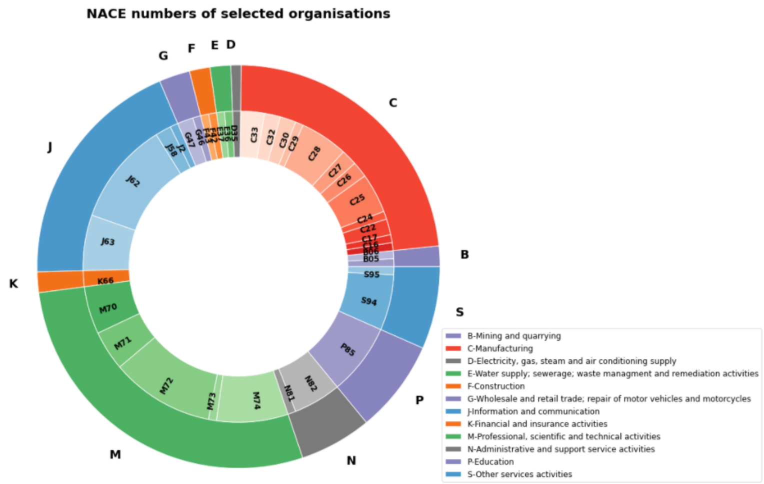


Figure 4
The distribution of the different pursued economic activities at present by the 51 Central European participants.

suggestions and due to its talkative output this technique had been selected to get a manageable picture on our textual inputs even for decision makers [14].

These findings from corresponding scientific literature suggest that topic modeling results shall be treated in a reserved way and used jointly with other methods and data sources in the presence of proper domain knowledge or compare the results with other topic modeling techniques like Correlated Topic Modeling or Embedded Topic Modeling [15], [16]. Nevertheless these techniques provide an automatized procedure for organizing, annotating and managing unstructured textual documents which would be hardly feasible manually it is still a question of debate how they shall be applied according to the research goals or how their results should be compared to each other (model checking problem) [17].

For text pre-processing open source python libraries were used. The *python NLTK 3.6.2.* toolbox was utilized for Natural Language Processing tasks, the *python wordcloud 1.8.1.* toolbox for text visualisation [18, 19]. Further text processing tasks for topic modeling were outlined using the *spaCy 3.0* and *gensim 4.0.1* python modules [20, 21]. The LDA topic modeling itself was carried out with the *gensim 4.0.1* module on the processed textual data and the results were visualised and the conclusions were drawn by using the *pyLDAvis 3.3.1.* module [22, 23].

During the LDA topic modeling the corpus of texts are mapped to a document - term matrix that encodes the occurrence frequencies of the words in each constituting document. Since texts typically serve with a skew distribution of words due to the various endings and synonyms this can lead to huge and sparse

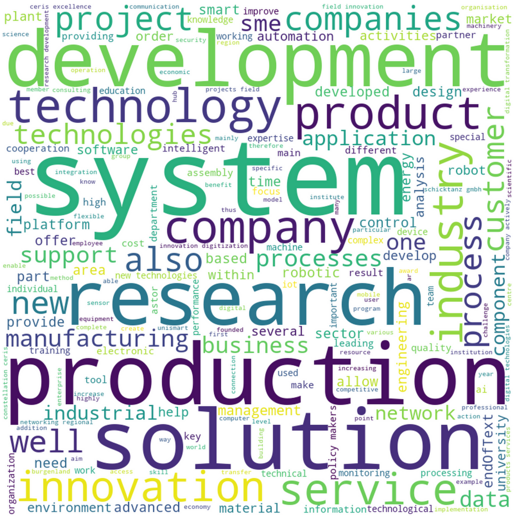


Figure 5
Exploratory data analysis of the available textual data after text pre-processing via *word cloud*.

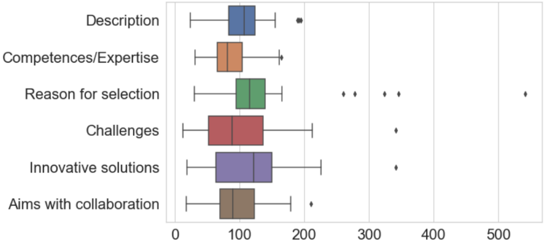


Figure 6
Exploratory data analysis of the available textual data.

document - term matrices that would negatively affect the topic model's predictive performance. Therefore, dimensionality reduction of data is essential via tokenization[6] and lemmatization[7], which can reduce variability of words while fairly keeping the content [8]. Further performance increase can be achieved by extending the vocabulary with bi- and tri-grams [8]. Such non - trivial bi- and tri-grams that occurred in our documents were were for instance: *smart_material*, *machine_learning*, *predictive_maintenance* and

[6] Splitting of text into list of sentences and sentences to list of words.

[7] Reduction of words to their corresponding vocabulary form, often by cutting off pre- and suffixes.

[8] Compound words formed from words that occur often together in text out of two- or three words. They are counted as one word in the document - term matrix in order to create more meaningful results by adding semantically complex terms.

developing_innovative_product. The followed topic modeling procedure is summarized on Fig. 7.

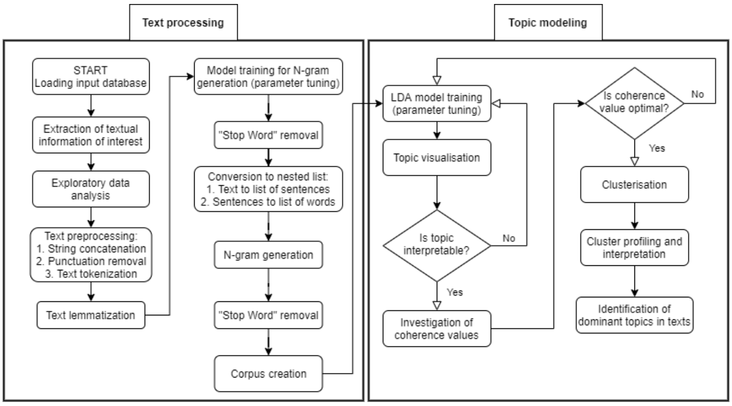

Text processing
Topic modeling
START
Loading input database
Extraction of textual information of interest
Exploratory data analysis
Text preprocessing:
1. String concatenation
2. Punctuation removal
3. Text tokenization
Text lemmatization
Model training for N-gram generation (parameter tuning)
"Stop Word" removal
Conversion to nested list:
1. Text to list of sentences
2. Sentences to list of words
N-gram generation
"Stop Word" removal
Corpus creation
LDA model training (parameter tuning)
Topic visualisation
Is topic interpretable?
No
Yes
Investigation of coherence values
Is coherence value optimal?
No
Yes
Clusterisation
Cluster profiling and interpretation
Identification of dominant topics in texts

Figure 7
Flowchart of the topic modeling procedure applied.

Since our exploratory data analysis provided the available text snippets to be relatively short and due to the nature of the described topics (see 2.1) are overlapping, synonymous and redundant text elements per organisation have been concatenated for each interviewee. Therefore 51 texts were analysed in our corpus throughout our topic modeling. A representation of the pre-processed texts of the corpus after dimensionality reduction is visualised in a *word cloud* form on Fig. 8. This summarising representation suggests that the organizations concerned were mainly interested in production and development related topics. Additionally minor interests regarding technology- and solution oriented issues could also be read off the figure.

For the selection of the number of topics the *C_v* topic coherence values have been investigated in an iterative parameter sweep. The topic coherence value serves with a simple topic score that gives the level of semantic similarity of high-scoring words in texts constituting the corpus. This quantitative metric is frequently used for topic number selection in case of LDA models for given hyper - parameters [24] [25]. A local (or global) maximum of topic coherence values were looked for with the topic numbers as independent variable. As it is indicated on Fig. 9 8 topics have been estimated as a reasonable local maximum resulted from the generated *C_v* scores.

On Fig. 10. the topics are viewed separately on an inter - topic distance plot that was generated with the *pyLDAvis 3.3.1.* python package. This impressive visualisation enables an easier interpretation of the topics at the selected optimum topic number. The size of the bubbles on the left are corresponding to the relative share of each topic in the corpus predicted by LDA. The horizontal bars on the right represent the frequency of the indicated terms[9] in the corpus

[9] Nomenclature for pre-processed words.

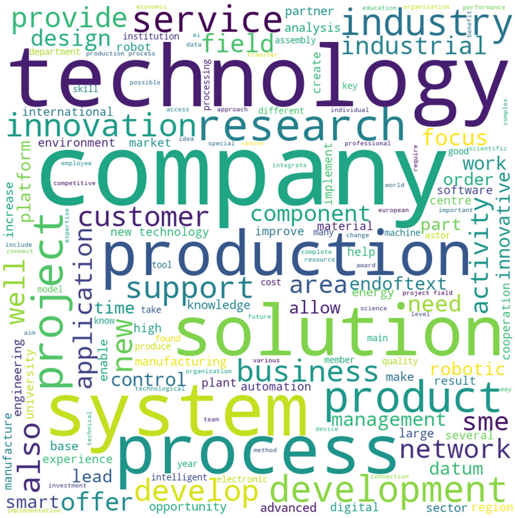


Figure 8

Word cloud representation of the resulted 51 texts after stemming and lemmatization.

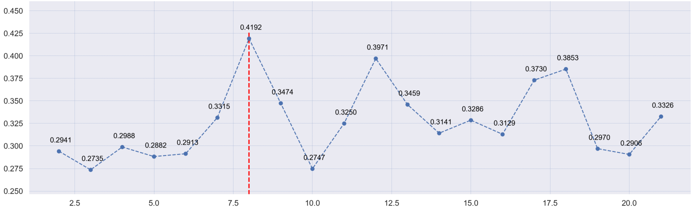


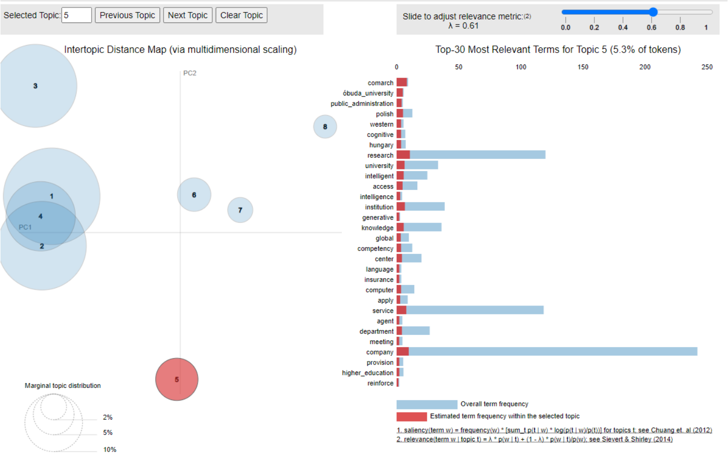
Figure 9

Selection of the optimal topic number based on the investigation of topic coherence values.

with blue and the estimated frequency of the same terms within the selected topic. More relevant terms for each topic (more topic specificity) are represented with longer red bars.

Based on this representation practitioners can assign meaning to the resulted topics and dominant topics can be determined to each partaker organization. This serves with a natural grouping possibility for the 51 business organization into 8 different groups. Table 4. summarizes the number of organizations assigned to each cluster and their relative weight within the sample. On the other hand Table 3. lists the most dominant keywords corresponding to each cluster withdrawn from *pyLDAvis* visualisation at $\lambda = 0.6$ term relevance factor that is suggested in [22].

The selection of most relevant keywords are somewhat subjective, since human decision has also to be involved to interpret a topic. In Table 3. three keywords



Selected Topic: 5

Slide to adjust relevance metric:(2)
λ = 0.61

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10
Visual representation of the resulted topics by the *pyLDAvis* python package.

Table 2
Number of organizations within the resulted clusters and their relative sizes compared to the whole sample.

Cluster ID	No. of members	No. members/Total (%)
Cluster0	16	31.4%
Cluster1	12	23.5%
Cluster2	11	21.5%
Cluster3	5	9.8%
Cluster4	3	5,9%
Cluster5	2	3,9%
Cluster6	1	2%
Cluster7	1	2%
Total:	51	100%

have been selected and the most descriptive and general term has been marked with bold in order to have a clear view on the resulted topics that are assigned to each cluster. As can be seen, around 88% of the total tokens could be assigned only to 4 topics, while the remaining 4 topics formed separated, smaller groups that can be regarded as separate outlying items (also see Fig. 10.).

The relationships between organizational types, NACE categories, countries of business organizations and the clusters provided by the LDA topic modeling have been visualised and investigated based on heatmaps and further interpreted (for the distribution of organisational types see Fig.11.).

The investigation of such heatmaps led to no correlation among any aforementioned variable and cluster indexes, however there was no cause to assume such either. Therefore, the clustering based on the given textual information offers a different grouping of members as the basic organizational information would suggest. This should be treated with precaution but as a further input for the investigation of our sample.

Table 3
Dominant keywords for each topic selected at $\lambda = 0.6$ term relevance factor to describe topics.

Cluster ID	Dominant keywords	Ratio of tokens (%)
Cluster0	**technology**, production, product	28.4%
Cluster1	**service**, system, development	23.9%
Cluster2	**customer**, company, solution	20.9%
Cluster3	**application**, robotic, engineering	14.7%
Cluster4	**intelligence**, university, competency	5.3%
Cluster5	**venture**, streaming, talent	3.4%
Cluster6	**furniture**, cluster, wood	1.8%
Cluster7	**engine**, worker, AR	1.6%

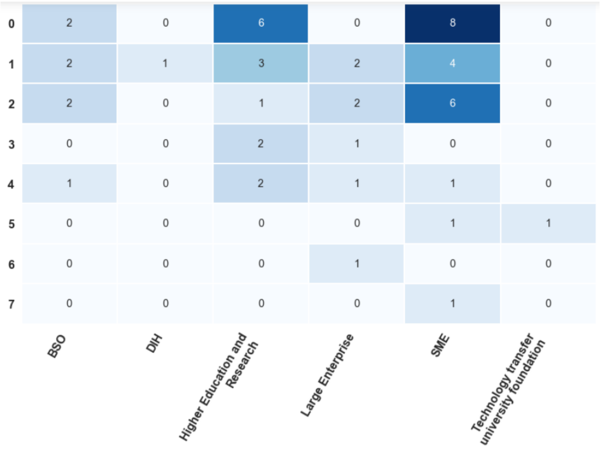


Figure 11
Number of partakers with different organisational types within the clusters suggested by *Latent Dirichlet Allocation* performed on the available textual data.

The presented method in this section can be regarded as a thorough and comprehensive overview on the available textual information in a way that is relatively easy to interpret. This can contribute to the understanding or determining of focuses and interests of business organizations and serve as a supplementary tool for policy makers in creation or promotion of more efficient and advantageous international cross-border cluster creations.

3.2 K-modes Clustering

Besides the investigation of general textual data available on the 51 Central European business organizations a more direct and company specific clustering has been performed based on the collected multiple selection choice questionnaire input. As it has already been mentioned in Section 2.1 this online questionnaire provided categorical data that do not fit regular clustering algorithms that expect numerical data on ratio scale[10] (e.g.: K-means).

[10] Type of numeric data on which the operation of division and a zero value are defined.

As our case shows there are many situations where categorical or mixed categorical and numerical variables are to be used for clustering purposes. The K-modes clustering algorithm is one of the most cited methods that enables partitioning data points characterized by such variables. The algorithm is based on the fundamental concept of the highly efficient K-means, thus it inherits many of its beneficial properties. It uses the same cost function as the K-means but a different dissimilarity measure as the Euclidean distance. According to literature results the K-modes needs less iteration to converge than the K-means, therefore its modification to categorical data results in a faster algorithm [10].

The dissimilarity measure applied by K-modes between two variables X and Y with respect to an attribute q_j can be defined as:

$$d(X,Y) = \sum_{i=1}^{n} \delta_{q_j}(x_i, y_i), \tag{1}$$

where

$$\delta(x_i, y_i) = \begin{cases} 0, \text{ if} x_i = y_i \\ 1 \text{ otherwise.} \end{cases} \tag{2}$$

This $d(.,.)$ dissimilarity measure forms a metric space on the set of categorical variables and the mode of a vector $\mathbf{X} = \{X_1, X_2, ...X_n\}$ that provides n observations and is described by m categorical attributes can be calculated as the minimum of [26]:

$$D(\mathbf{X}, Q) = \sum_{i=1}^{n} d(X_i, Q) = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_{i,j}, q_j) \tag{3}$$

The steps of the algorithm given by Huang in [10] can be summarised as follows:

1. Initialisation of cluster centres by selecting K modes
2. Object allocation to the nearest cluster modes according to the distance metric of Eq. 1. Update of the modes of the clusters based on the minimum value of Eq. 3.
3. Reallocation of objects to the modes in case the updated modes resulted closer than the previously allocated ones.
4. Iterative repetition of the previous step until no object changes its assigned cluster anymore.

As K-modes inherits many features of K-means thus serves at least a locally optimal solution. This is however dependent on the number of categories and the initialisation of the algorithm [10], [26]. The initialization is a crucial part of the calculations since there is no guarantee to find a global minimum. On the other hand the likelihood of finding outlying data points for cluster centers and failing to describe the bulk of the data is relatively high. In such cases (if outlier detection belongs not to our goals) the data structure cannot be revealed with a fixed cluster number [27]. To alleviate this issue various techniques are listed in literature that can help in case of large, multidimensional categorical data sets for instance by introducing partitional entropy, density measures combined by identification of representative elements or using multiple clustering [27], [28] [29–31].

In our case the *Cao* initialisation has been chosen that takes also into consideration the density of data besides the distance metric defined in Eq. 1. This initialisation was especially suitable because a small data set had to be partitioned. Having performed several clustering with different initial seed numbers 5 clusters were selected as an optimum number. Nevertheless, in the vicinity of the selected number, other clustering results were investigated regarding interpretability of the results. At this point experts were involved who concluded that groupings with different cluster numbers than 5 served with less meaningful partitioning.

Figure 13. represents the distribution of preference choices of thematic fields for a possible future co-operation selected by the attendees. Since each partaker could select more fields of interest the summarizing heat-map represents the frequency of a given thematic field selected by the partakers in each cluster. Thus neither the columns nor the rows add up to 100%.

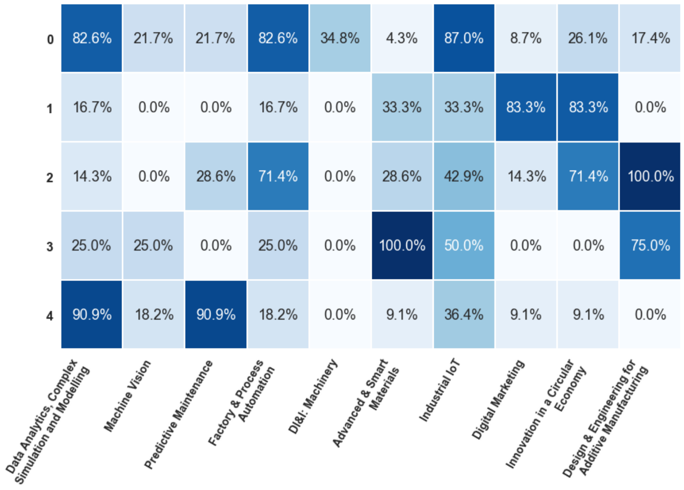


Figure 12
Relative distribution of preferences for co-operation within the resulted 5 clusters provided by the K-modes algorithm.

The interpretation of the resulted clustering based on expert opinions according to Figure 13. is the following:

- **Cluster0:** Group of participants interested in Industry 4.0 related technologies with special emphasis on *Data Analytics, Factory&Process Automation* and *Industrial IoT*.

Table 4
Number of members within each cluster provided by the K-modes algorithm and their relative size compared to the sample size.

Cluster ID	No. of members	No. members/Total (%)
Cluster0	23	45,1%
Cluster1	6	11,7%
Cluster2	7	13,7%
Cluster3	4	8%
Cluster4	11	21,5%
Total:	51	100%

- **Cluster1:** Group of participants interested in Marketing and Sustainability related fields with special emphasis on *Digital Marketing* and *Innovation in a Circular Economy*.
- **Cluster2:** Group of participants interested in Sustainable production and Automation with special emphasis on *Factory & Process Automation, Innovation in a Circular Economy* and *Design & Engineering for Additive Manufacturing*.
- **Cluster3:** Group of participants interested in material science related technologies with special emphasis on *Advanced & Smart Materials* and *Design & Engineering for Additive Manufacturing*.
- **Cluster4:** Group of participants interested in data science related technologies with special emphasis on *Data Analytics, Complex Simulation and Modeling, Predictive Maintenance* and *Industrial IoT*.

In Table 4. the number of partakers are listed that were assigned to each cluster by the K-modes algorithm. Within the 5 groups *Cluster0* became the largest with 23 members. This constitutes little less than half of the total sample. Having investigated the heat-map visualisation on Fig. 13. a relatively diverse set is revealed where besides SMEs or Large Enterprises, Higher Education and Research Institutions are also concerned at approximately the same proportion.

Further inferences that can also be depicted from the relative distribution of the organisational types within the resulted clustering is:

- Business Support Organisations (BSO) were mainly interested in marketing related activities and were therefore highly represented in *Cluster1*, which is in accordance with their fundamental business orientation.
- SMEs were mainly interested in technology related fields and novelties. Thus their presence in *Cluster1* was less prominent, however the remaining 4 clusters contained SMEs in a relatively high share. SMEs appeared to be widely interested in several fields and sought opportunities for innovation and collaboration in most of the offered options listed in the survey.
- Large Enterprises were mainly interested in automation and optimization in order to gain bigger or new markets. Therefore they represented

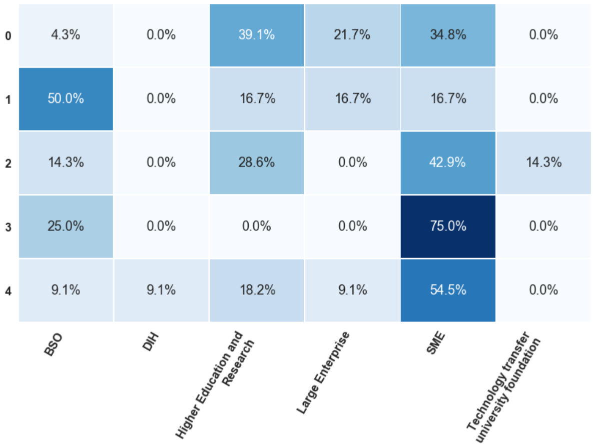


Figure 13
Relative distribution of organisational types within each cluster provided by K-modes.

themselves in *Cluster0* and *Cluster4* regarding technological interest and *Cluster1* regarding marketing activities. Most probably due to their fixed, rigid and mature production system, they were more restricted by their circumstances. It can be assumed that they have utter individual goals and do not intend to invest time and energy in other innovative fields apart from their focus (e.g.: *Design & Engineering for Additive Manufacturing* or *Circular Economy*).

The partitioning of the 51 Central European business organization that took part in our online survey, based on their multiple preference choices for a possible future co-operation and further education is also visualised on Fig. 14. on a map layout. This geographical visualisation shows that with the present method a non-trivial grouping of the members could be achieved in an algorithmic way with a well-defined methodology that served with a possible cross-border clustering recommendation for the partakers.

As the presented partitioning indicates, the output provided by the K-modes algorithm can be logically interpreted. Since our sample included partakers from different industries, Technology Transfer- and Education Institutions furthermore supporting organisation as well, according to literature the establishment of such cross-border partnerships that can operate more innovative has a higher probability. Moreover, the clusters formed well-definable interests that could be assigned as a label to them during the profiling process. For the comparison of the two partitioning – resulted by LDA for textual data and K-modes for categorical data – the *Rand Index* has been used [32]. On a ±1 scale. The calculated *Rand Index* resulted to be only -0.007, which indicates that the two data sets with different methodology provided completely different aspects for the grouping of the 51 members. Hence, the advise of the authors is that the results of the LDA topic modeling

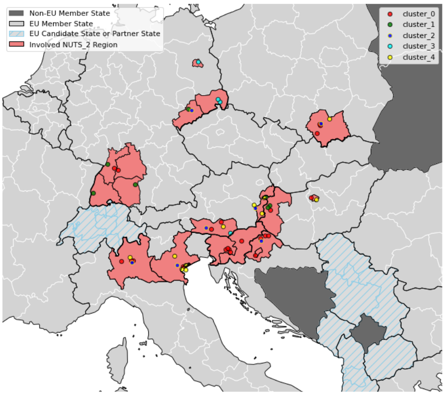


Figure 14
Geographical distribution of the resulted clusters provided by the K-modes algorithm.

shall be treated as an additional step of exploratory data analysis and the K-modes clustering as the basis for partitioning the organizations into partnerships.

Conclusions

In our present study textual information and multiple selection choice feedback collected in an online survey have been analysed in order to facilitate current selection and grouping processes of business organisations for future co-operations. The data originated from 51 Central European stakeholders which from organizational point of view were business-, education- and research organisations or institutions with various industrial or scientific focus areas and experience.

In the form of the multiple selection choice survey categorical variables were received on the preferences of the partakers on possible up-to-date industrial- and scientific fields. These could be possible areas for a future collaboration among the selected members may it be a profit oriented or R&D activity in short- or even in long-term. A K-modes based clustering has been performed on the processed data and a well-interpretable grouping of the organizations has been outlined.

The results of the provided clustering has been compared with additional geographical, industrial branch and descriptive textual information collected to get a complex, comprehensive picture on the organizations and their proposed grouping. The investigated textual data incorporated short descriptions on the organisations' focuses, main operation areas, activities, challenges and achievements that was used as a basis for a topic modeling approach. Our choice for an exploratory analysis fell on *Latent Dirichlet Allocation* (LDA), which is one of the most wide-spread topic modeling methods used by practitioners and researchers.

Having investigated the resulted clusters in terms of geographical location, organisational type and industrial branch it can be concluded that a non-trivial but an easy to interpret categorisation of the partakers has been achieved. This combined with the exploratory data analysis and topic modeling on easily accessible, company related further data can extend the knowledge (identification of dominant topics, keywords etc.) on the sample group of interest. In the end of the day this can contribute to a more efficient networking of companies and institutions not just on regional but also on international level by not just taking industrial or business orientation into account and can save a huge amount of energy and resources by connecting the right stakeholder for a partnership establishment. Our hope is that the conducted procedure can provide a useful tool in the long run for business actors and decision makers to promote cross-border international partnership forming and thereby contribute to the increase in competitiveness of regions on the European Union level.

In a future work we would like to extend our present findings by incorporating other topic modeling techniques like *Correlated Topic Modeling* or *Embedded Topic Modeling* that can also take correlations among resulted topics into account or can handle short texts and documents with highly skewed vocabulary more efficiently and more robust. This could illustrate textual content preferably and further support and complement our results gained from clustering thriving.

Acknowledgment

Gy. Eigner was supported by the Eötvös Loránd Research Network Secretariat under grant agreement no. ELKH KÖ-40/2020 ('Development of cyber-medical systems based on AI and hybrid cloud methods'). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 679681). Project no. 2019-1.3.1-KK-2019-00007. has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme. The publication of this article has been supported by the Robotics Special College via the "NTP-SZKOLL-20-0043 Talent management and community building in the Robotics Special College of Óbuda University" project and by the Pannon Business Network Association. The publication was supported by the Applied Informatics and Applied Mathematics Doctoral School of Óbuda University.

References

[1] A. Reveiu and M. Dardala. The Role of Universities in Innovative Regional Clusters. Empirical Evidence from Romania. *Procedia - Social and Behavioral Sciences*, 93:555–559, 2013.

[2] A. Némethné Gál. Competitiveness of Small and Medium-sized Enterprises - PhD Thesis (in Hungarian). 2009. István Széchenyi University, Doctoral School of Regional- and Business Administration, Győr.

[3] Y. Kajikawa, J. Mori, and I. Sakata. Identifying and Bridging Networks in Negional Clusters. *Technological Forecasting and Social Change*, 79(2):252–262, 2012.

[4] Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima. Multiscale Analysis of Interfirm Networks in Regional Clusters. *Technovation*, 30(3):168–180, 2010.

[5] C. Felzensztein, E. Gimmon, and K. R. Deans. Assessment of Safety Performance in Indian Industries using Fuzzy Approach. *Industrial Marketing Management*, 69:116–124, 2018.

[6] Computer software. Vers. 2-2.4.0. Anaconda. Anaconda Software Distribution. *Web.: https://anaconda.com*, Nov. 2016.

[7] K. Jordahl. Geopandas: Python Tools for Geographic Data. *Web.: https://github.com/geopandas/geopandas*, 2014.

[8] K. Bastani, H. Namavari, and J. Shaffer. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints. *Expert Systems with Applications*, 127(1):256–271, 2019.

[9] S. Daenekindt and J. Huisman. Mapping the Scattered Field of Research on Higher Education. a Correlated Topic Model of 17,000 Articles, 1991–2018. *Higher Education*, 80(3):571–587, 2020.

[10] Z. Huang. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[11] F. Cao , J. Liang, D. Li, L. Bai, and C. Dang. A Dissimilarity Measure for the k-Modes Clustering Algorithm. *Knowledge-Based Systems*, 26(3):120–127, 2012.

[12] G. Li, X. Zhu, J. Wang, D. Wu, and J. Li. Using LDA Model to Quantify and Visualize Textual Financial Stability Report. *Procedia Computer Science*, 122:370–376, 2017.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[14] I. Vayansky and S. A. Kumar. A Review of Topic Modeling Methods. *Information Systems*, 94(101582), 2020.

[15] D. M. Blei and J. D. Lafferty. Correlated Topic Models. *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 147–154, 2005.

[16] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8(2):439–453, 2020.

[17] G. Beriha , B. Patnaik, S. Mahapatra, and S. Padhee. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84, 2012.

[18] Bird, Steven, Edward Loper and Ewan Klein. Natural Language Processing with Python. *O'Reilly Media Inc.*, 2009.

[19] A. Mueller. Python WordCloud Documentation. *Web.: https://amueller.github.io/word_cloud/*, 2020.

[20] spaCy Python Module. https://spacy.io/usage/linguistic-features. Accessed: 2021-06-02.

[21] R. Rehurek and P. Sojka. Gensim–Python Framework for Vector Space Modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[22] C. Sievert and K. E. Shirley. Ldavis: A Method for Visualizing and Interpreting Topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Jun. 2014.

[23] pyldavis Python Module. https://github.com/bmabey/pyLDAvis. Accessed: 2021-06-02.

[24] S. Rani and M. Kumar. Topic Modeling and its Applications in Materials Science and Engineering. *Materials Today: Proceedings*, 45(6):5591–5596, 2021.

[25] V. Gangadharan and D. Gupta. Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques. *Procedia Computer Science*, 171:1337–1345, 2020.

[26] F. Cao, J. Liang, and L. Bai. A New Initialization Method for Categorical Data Clustering. *Expert Systems with Applications*, 36(7):10223–10228, 2009.

[27] F. Jiang, G. Liu, J. Du, and Y. Sui. Initialization of K-Modes Clustering Using Outlier Detection Techniques. *Information Sciences*, 332(1):167–183, 2016.

[28] L. Bai, J. Liang, and C. Dang. An Initialization Method to Simultaneously Find initial Cluster Centers and the Number of Clusters for Clustering Categorical Data. *Knowledge-Based Systems*, 24(6):785–795, 2011.

[29] G. Beriha , B. Patnaik, S. Mahapatra, and S. Padhee. Cluster Center Initialization Algorithm for K-Modes Clustering. *Expert Systems with Applications*, 40(18):7444–7456, 2013.

[30] T. Stepišnik, D. Kocev, and S. Džeroski. Option Predictive Clustering Trees for Multi-label Classification. *Acta Polytechnica Hungarica*, 17(10), 2020.

[31] A. S. Kazsoki and B. Hartmann. Hierarchical Agglomerative Clustering of Selected Hungarian Medium Voltage Distribution Networks. *Acta Polytechnica Hungarica*, 17(4), 2020.

[32] L. Hubert. Comparing Partitions. *Journal of Classification*, 2:193–208, 1985.