# Literature insights on Economic Complexity with Text Mining

BENNOUR Mohamed Hsin – HGXGPE

2024-05-14

## Abstract

In this paper I decompose the literature body of the Economic Complexity paradigm using text mining tools. The data used is a set of 1057 abstracts extracted from the publically available sources, I describe the selection criteria of the data, the preprocessing phase, the modeling and testing. Eventually I provide a an oiverall interpretation of the results with the necessary nuance with such an approach. The modeling phase resulted in 4 topics that encapsulates the 4 major elements of the literature, although with this same approach using full papers instead and a better refinment of the selection criteria of the corpora can yield much more detailed and insightful results.

## Introduction

Knowledge comes in multiple forms, we can observe knowledge in technical artifact and gadgets (physical), we can also refer to it in books and documentation, as well as tacit knowledge that is formed out of experience and exposure and is not physically encoded and is hard to spread (Balland et al. 2022). Measuring the contribution of tacit knowledge has been alaways problematic for researchers as there's no certain way to properly aggregate it into an indicator that dignifies it and conserve the information. The paradigm of Economic Complexity (or EC) treats this problem by using different aproaches from network science in order to captures the interaction between different agent and conserve the information within a specific system enabling an appropriate approximation of the contribution of such tacit knowledge to that system. In this context EC is considered as tool box used by various researchers in different areas of scientific research. However, given the tremendous size of publications in this context, we cannot be sure about the major trends and lines of research that this paradigm involves, in fact (Hidalgo 2023) provided an interesting approach in his analysis of the current state of the literature in EC and potential future directions of research. From this perspective, summarising the content of such a literature body might also contribute to a more targeted research and eventually to a better understanding of the potential of such a paradigm. To do this, one might go beyond a systematic literature review, and look for ways to automatise the search and classification of this body of literature that started with the early works of (Hidalgo et al. 2007). To achieve such a classification, we will need to resort to sophisticated tools and adopt non-conventional approaches of analysis.

In this context, text mining is a powerful tool that can be used for various objectives. The power of such tool has been seen in 2020 when researchers all over the world needed information and details of studies in many medicine-related fields for their endeavor in developing a vaccine for covid-19. Additionally, the emergence of large language models (or LLMs) is another realisation taken to the extreme of text mining and language modeling. In academia, researchers can also leverage such tools to inform their decisions in pursuing a research field for instance. Eventually everything, including text, boils down to data and with the proper tools we can analyse it and find insightful information. In this context, this paper aims at achieving just that. In fact, I will be using text mining techniques to delve into the Economic complexity literature as it relates directly to the topic of my thesis. The idea here is to cluster a set of publications' abstracts with a topic modeling technique (namely the Latent Dirichlet Allocation (LDA) algorithm, (Blei, Ng, and Jordan 2003)). The idea is to explore this body of literature beyond the keywords and the subject categories (although useful to consider), by exploring analytically the relevance and significance of the words and terms used in a set of a 1057 abstracts extracted from the Web of Science (WoS) database, and combined into

a single text corpus(more details in the following sections). The justification for such an approach is three folds. First, the economic complexity literature gained a tremendous momentum since its inception in 2007. Second, the problems treated in this body of literature are very diverse and the contributions are various and marginal in many cases, thus by having more context from the abstracts I can identify hidden areas of research that are not explicitly treated. And third, it's very interesting to see the trends around the topics, and what are the current main concerns of these studies. Another aspect of the importance of such an approach is the fact the Economic Complexity paradigm offers a set of network analysis tools that can be leveraged in any sense and direction (Hidalgo 2021). For instance there are studies that unravel the relation of complex networks and emissions (Romero and Gramkow 2021) and others with gdp growth (Koch 2021 ; Chávez, Mosqueda, and Gómez-Zaldívar 2017). Thus to investigate the trends and the directions of this literature body, in bulk and without the biasis that might be find in reviews of literature, text mining provides the best alternative. The remaining of the paper is as follows, in section 2 I present the data in more details and the preprocessing that took place to become suitable for the analysis which I will uncover and detail in section 3. In section 4 I will present the main results and interpret them. In section 4 I will provide visualisations for the results and discuss them, and finally I conclude the paper in section 5.

## Data preprocessing

The data used in this paper was extracted from the WoS (Web of Science) databse, the extraction process is quite simple and is done directly from the web interface of the website after a search query that targets papers investigating complexity and relatedness using the following :

```
# Topic
TS=("economic complexity" OR "relatedness")  AND
# Author Keywords
AK =("economic complexity" OR "relatedness")
```

The result of this search yielded 1057 papers in total after limiting the years to the range between 2007 and 2024 and the WoS subject category to subjects in economics, regional and urban planning as well as geography of innovation, I then export the metadata of these papers (Author, Title and Abstract) into multiple tab delimited files with `.txt` extension. The aim of this work is to target the complexity literature through the texts of the abstracts of the extracted data. Thus I merge all the abstract into one corpus and start the preprocessing from that point.

The preprocessing starts with the few typical tasks, I first lower all cases in the corpus text, I removed all the non english letters and extra spaces, as well as English stopwords as well as other words that might be over used the likes of (use, also, one, two, etc.). Additionally I proceed with the removing numbers and punctuation.

The second step of the preprocessing envolved the tokenisation of the corpus, this process usually is followed by the stemming procedure, but judging from the small data size and from the experiments I conducted in the process, I decided to avoid that. With these steps, I finalised the preprocessing phase with the creation of the term matrix.

## Generating topic models

For the topic modelling phase, I used the Latent Dirichlet Allocation (LDA) algorithm (Blei, Ng, and Jordan 2003). For this algorithm many methods for sammpling and estimations has been developed such as the Gibbs sampling method (Porteous et al. 2008) and VEM (Variational Expectation-Maximization) method (Nasios and Bors 2006). The Gibbs sampling method is a Markov Chain Monte Carlo procedure that is used with LDA to estimate the posterior distribution of the hidden values, the topics in our case. This method is quite efficient for small datasets such as the one used here wheras the VEM method uses variational inference to estimate the posterior distribution as it transforms the inference problem into an optimization problem, The VEM method is more reliable when dealing with huge corpus of texts, however it provides an approximation to the posterior probabilities making it faster than the Gibbs method but at the same time less accurate. Thus for the purposes of this paper I will use the Gibbs method of sampling as it aligns

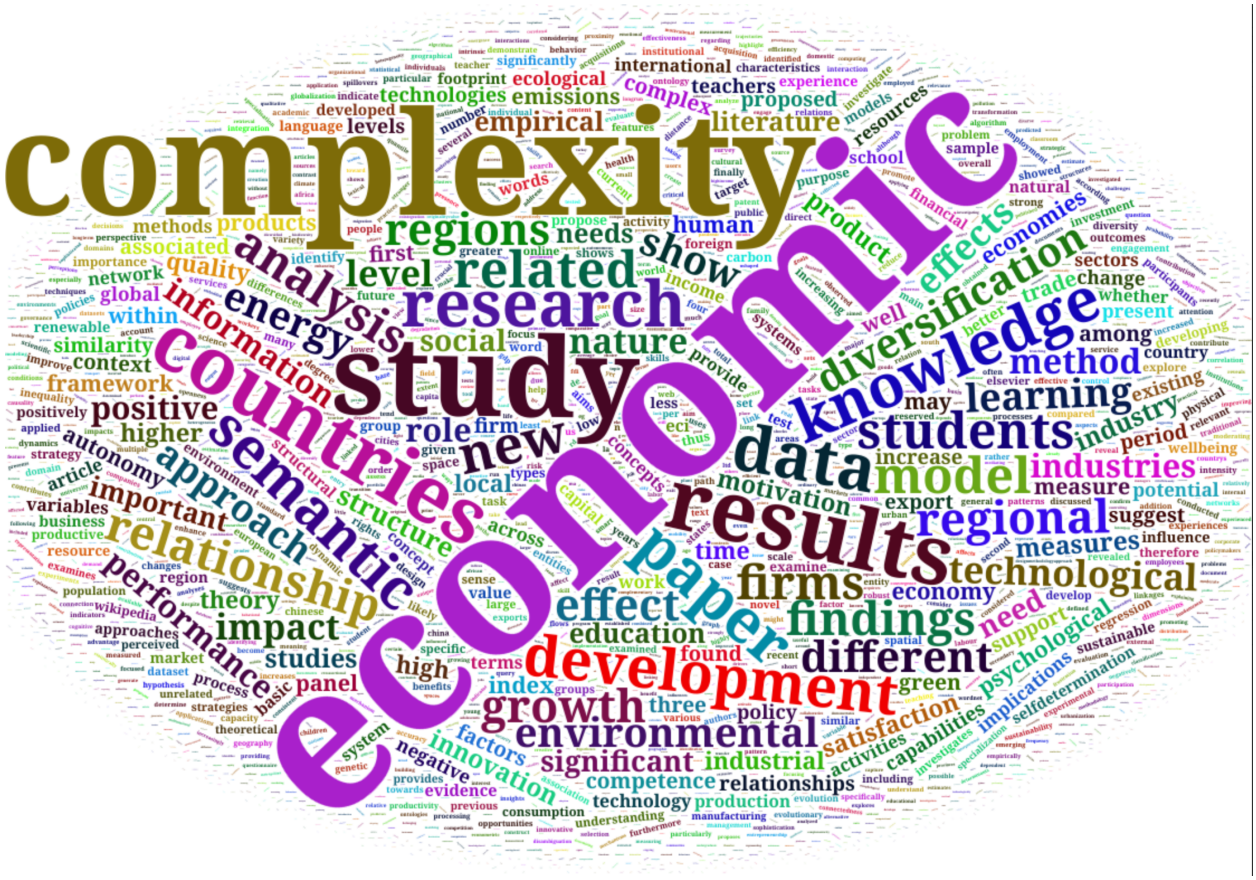



Figure 1: Word Cloud of the corpus

with the data I have. However, I should note that this method is more effective when is conducted to model multiple documents rather than one corpus such as the case in this paper.

Additionally for this phase, I created a grid of parameters that controls the model"s accuracy to model the topics in the corpus:

- Alpha ($\alpha$): is a hyperparameter that influences the distribution of topics within documents. It is a parameter of the Dirichlet prior on the per-document topic distributions.
- Iter (Iterations): The number of iterations (iter) is a parameter that specifies how many times the Gibbs sampling process should be repeated.
- K (topics): The number of topics to model for each model variation

To optimize the Latent Dirichlet Allocation (LDA) model for topic modeling, a comprehensive set of parameter configurations was tested using Gibbs sampling. The parameters included different values for alpha, the Dirichlet prior for document-topic distribution, ranging from 0.01 to 5, specifically: 0.01, 0.05, 0.1, 0.5, 1, 1.5, 2, and 5. Additionally, the number of iterations for the Gibbs sampler was varied extensively, testing values of 100, 200, 300, 400, 500, 1000, and 2000 iterations to ensure convergence and stability of the results. As for the number of topics k, was initially varied to explore different insights of topic decomposition, with values of 3, 4, 5, 6, and 7 topics being tested. This gets us eventually 280 combinations all together of these parameters which is quite the process to evaluate manually, although eventually I used the Arun2010 metric (Arun et al. 2010) (also known as the The Kullback-Leibler divergence) to determine the best number of topics I can have given the term matrix, and 4 ended up being the most optimal number of topics for the corpus. Additionally, and to ease up the evaluation of the models, I rerun the models again with 4 topics and the same grid of values for the parameters alpha and number of iterations, and calculated the perplexity measure for these models (yielding 56 combinations) to look at the most coherent ones at least mathematically before evaluating the content of the topics themselves. In this case, perplexity-as described in (Neishabouri and Desmarais 2020)-is one of the most popular metrics used in the text mining and language modeling. This metric is based on estimating the probabilities of an unseen test data that's normalised in order to evaluate the overall goodness of fit of these models. Finally, The top 10 models with the lowest perplexity measure are shown in table 1.

Table 1: Models' perpelexity (best 10 models)

| | Perplexity |
|---|---|
| alpha_2_iter_1000 | 2636.806 |
| alpha_0.5_iter_2000 | 2636.824 |
| alpha_0.1_iter_2000 | 2636.836 |
| alpha_0.01_iter_2000 | 2636.876 |
| alpha_2_iter_2000 | 2636.881 |
| alpha_0.01_iter_1000 | 2636.938 |
| alpha_1_iter_2000 | 2636.942 |
| alpha_5_iter_2000 | 2636.973 |
| alpha_1.5_iter_2000 | 2637.013 |
| alpha_5_iter_1000 | 2637.025 |

Additionally, Table 2 provides a glimpse on the terms in each topic for the top 3 models with the least perplexity values. Each model provides 4 topics with the top 7 terms each.

Table 2: Best 3 models (topics and terms)

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| alpha_2_iter_1000 | | | |

| | | | |
|---|---|---|---|
| gdp | individual | n | economic |
| part | proposes | practices | relatedness |
| entry | performed | focusing | complexity |
| representation | year | distribution | study |
| mechanism | revealed | probability | results |
| density | contexts | way | countries |
| cluster | patents | reveals | research |
| alpha_0.5_iter_2000 | | | |
| question | economic | evolutionary | provide |
| ma | relatedness | conditions | long |
| practices | complexity | part | shown |
| employee | study | creation | investigated |
| alternative | results | finding | classification |
| proposes | countries | policymakers | qualitative |
| gap | research | frequency | primary |
| alpha_0.1_iter_2000 | | | |
| estimate | common | economic | often |
| selection | improving | relatedness | connection |
| urbanization | job | complexity | motivational |
| vector | relationships | study | creative |
| articles | sport | results | algorithms |
| organizational | properties | countries | longrun |
| capture | collected | data | transition |

Given the small number of observations that I have, and potentially some issues I might not have caught with the preprocessing, I decided with proceeding with this approach. The next step will involve interpreting the results from the top 3 models with the least perplexity similarly to (Griffiths and Steyvers 2004).

## Topic identification

The modeled topics in table 2 provide directional insights on the literature on Economic Complexity. From the results obtained in table one (and detailed in figure 1), I decided to adopt the topics from the first model (alpha_2_iter_1000). As indicated by the model name, the parameters are $\alpha = 2$ and the number of iterations is 1000. The reason for this choice is, apart from having the lowest perplexity of all the other models, it also provides the most diverse topics and meaningful terms that actually might align with the literature. For instance the first topic (with the terms: gdp, part, entry, representation, mechanism, density, cluster) can be interpreted as the body of this literature that targets the relatedness density of clusters of agents (institutions, regions, countries, etc.) how this density is affected by the entery to a new area of production or knowledge, and essentially explaining the gdp changes by those dynamics, a major seed paper in this context is (Boschma 2017). This topic can be called *relatedness* which is one major component of the Economic Complexity paradigm and its literature. The second topic (containing terms : individual, proposes, performed, year, revealed, contexts, patents) is describing the contexts of the specialisation patterns (revealed as in Revealed Comparative Advantage) and the dynamics of these specialisation (over time: year) by the means of patents' applications data. This topic describes the mechanics of specialisations of innovative firms and thus can be named the *innovation* topic, which is one of the most treated areas in this paradigm as can be seen in seminal works such as (Balland et al. 2018 ; Balland and Rigby 2017). Moreover, the third topic (containing: n, practices, focusing, distribution, probability, way, reveals) seems to target the statistical and mathematical sides of the toolkit provided by the paradigm. Indeed, many models have been developed in this body of literature to assess the probability of entery to a new field of knowledge given specific variables, such as the level of relatedness density, complexity, and in other instances even gdp. In fact some leading scholars in this body of literature argued continuously about the need to develop new methodologies that allow researchers to properly investigate the contribution of different factors in increasing the likelihood of

entery to a new field in a more dynamic sense, studies like (Tacchella et al. 2012 ; Broekel 2019 ; Guerzoni, Nuccio, and Tamagni 2024) have answered this call and contributed technically to the literature. Being able to determine the optimal diversification and specialisation strategy (over time and space) given the corpus of the knowledge that a study captures is a powerful concept and has been lately addressed in many papers most notably in (Alshamsi, Pinheiro, and Hidalgo 2018). Thus this third topic will be named the *technical* cluster. Finally, the forth topic (with the terms: economic, relatedness, complexity, study, results, countries, research) represent the classic cluster of this literature focusing mainly on explaining economic performance (in terms of GDP, or attractiveness for FDI and trade (Sadeghi et al. 2020) or in many instances emission and energy consumption (Abbasi et al. 2021 ; Kazemzadeh et al. 2023)) with the characteristics of a country in terms of its complexity (the level of diverse specialisations) and relatedness density (the diversity of related knowledge in which a country has a comparative advantage). Thus this topic should be called the *complexity* topic, and can be observed in many studies contributing to the way Complexity measures are computed such as (Cristelli et al. 2013 ; Ivanova et al. 2017 ; Ivanova, Smorodinskaya, and Leydesdorff 2020).

## Visual exploration

In this section I will interpret with a bit of nuance the wights of each term in each topic of the previously modeled corpus. In fact a summary of these weights can be found in Figure 2. We can see from the figure that topic 4 (*Complexity*) contains the terms with the highest weights across all the 4 topics. This is understandable given that Economic Complexity as a literature body is fairly new and that more contributions in the core of the literature is being done. Although the term "economic" seems to have the greatest weight of all, which also makes perfect sense since first, the term is part of the naming of the paradigm itself, and second this paradigm mainly provides tools that target the analysis of economic activity in general. In the same context, for topic 1 (the relatedness topic) the terms gdp, part, and entery are the highest defining terms of this cluster meaning that the economic performance and entery to new areas of knowledge are of major concern in this topic, although I cannot relate the relevance of the term "part" in this topic to anything in the literature. Similarly with topic 2 ("Innovation") the highest contributing term to this cluster is "individual", a term that is not necessarily meaningful in contrast with the other terms, but can be considered as proxy to the individualistic aspect to an innovative activity, this can also be justified by the appearance of the term "patents"-although with the least weight- the combination can be made sense of. The idea here is not to make sense of the weight of each word in each topic, but rather the collective meaning for all the terms in a specific topic instead. Finally, for the third topicd "technical", we can see that the term "n" has the highest weight, which to some extend logical as technical papers tend to identify parameters even in the abstract. Additionally the terms "probability" and "distribution" appear to have the same weights which can explain the popularity of probabilistic approaches in this body of literature.

Figure 2: Prior weights of each words in each topic

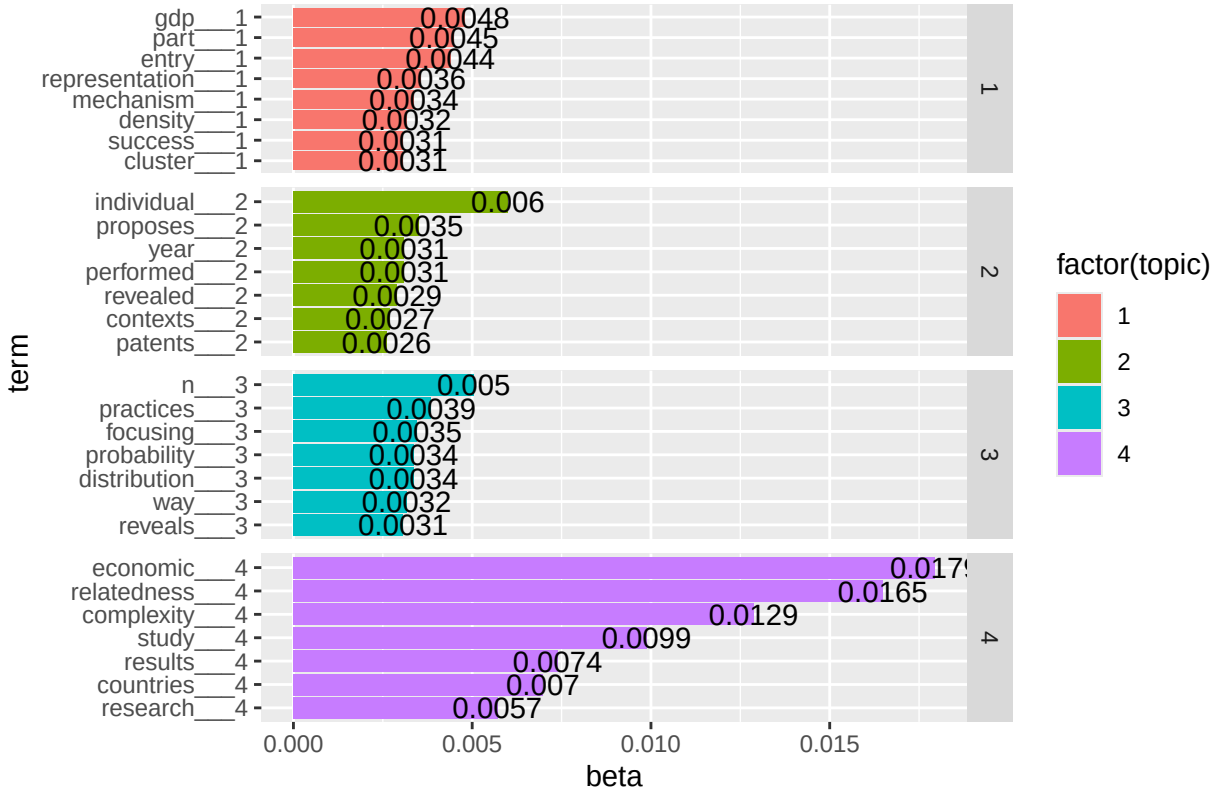
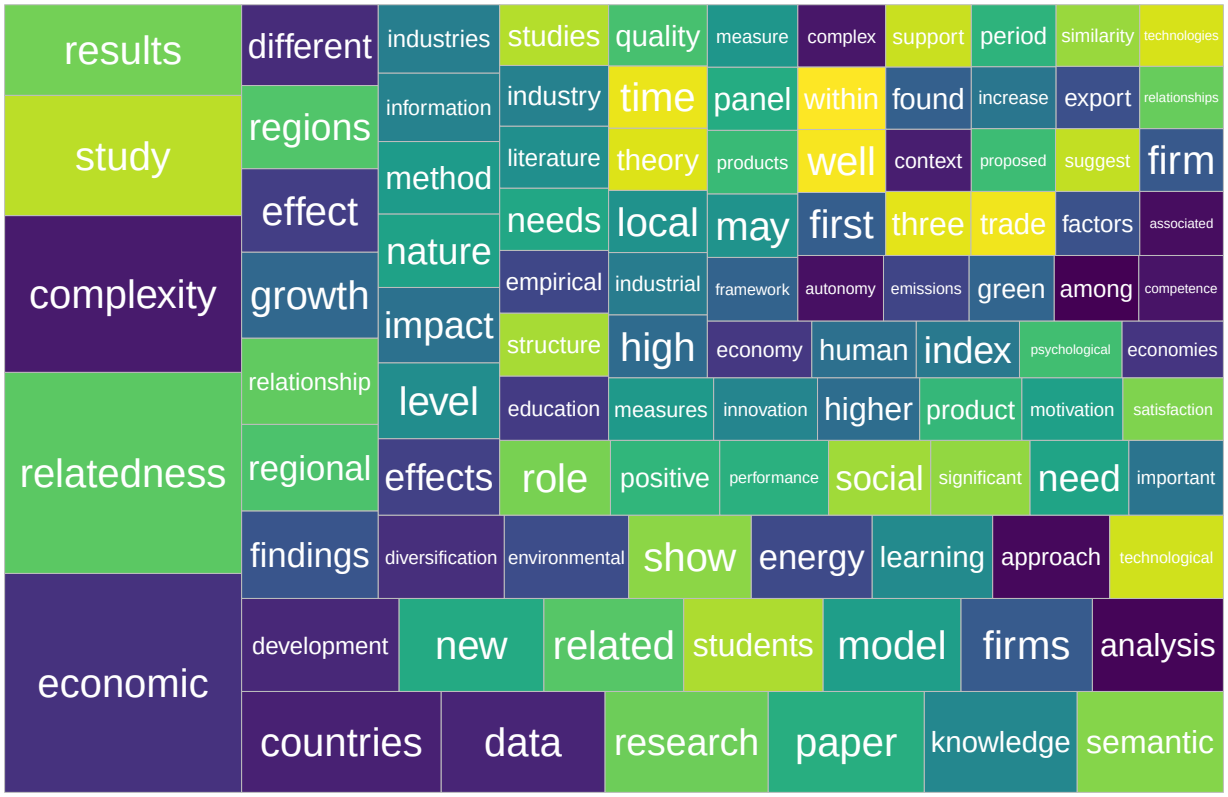


Figure 3: Treemap of the corpus




## Conclusions

In conclusion, this work is a mere trial to model topics of a sample of abstracts of the body of literature in Economic Complexity. In this paper, I used preprocessing techniques to create a single corpus that incorporates all the abstracts, then I used LDA with different combinations of parameters, from which I chose the model with the parameters yielding the lowest perplexity measure. The results that this modeling gave are mixed, on one hand the forth topic ("Complexity") provides a general and meaningful outlook with the classic terms that can be found in the papers of this body of literature. However for the other topics, and despite my effort to initially bend the terms and interpret their meaning in the context of the literature itself, the weights associated with each of these terms show unconsistencies and fuzziness that cannot be properly interpreted. This can be due to various factors, One usually in the abstracts the authors try to justify the study and present the methodology and results in a brief manner, and thus the abstracts might not properly capture the terms that are used across the full papers. Second, the preprocessing can be improved especially regarding the stemming and the stop words. Third, other options of modeling algorithms should be considered in order to improve the results and their interpretability. Finally, and probably most importantly, more data is needed, the size of 1000 publication can be decent if it contains the entirety of the publication and not just the abstracts. Eventually, this excercice was an interesting application to topic modeling on a body of literature I'm working with for my thesis. Initially the idea was to try and compare WoS subject categories with the topics, although the initial data contains 35000 records of published studies from 152 category, my intuition was that it would be too much to model such corpus. After this paper, the next steps would be to actually test the same approach but on different corpora that represent unique subject categories.

## References


Abbasi, Kashif Raza, Kangjuan Lv, Magdalena Radulescu, and Pervez Ahmed Shaikh. 2021. "Economic Complexity, Tourism, Energy Prices, and Environmental Degradation in the Top Economic Complexity Countries: Fresh Panel Evidence." *Environmental Science and Pollution Research* 28: 68717–31.

Alshamsi, Aamena, Flávio L Pinheiro, and Cesar A Hidalgo. 2018. "Optimal Diversification Strategies in the Networks of Related Products and of Related Research Areas." *Nature Communications* 9 (1): 1328.

Arun, Rajkumar, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. 2010. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations." In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part i 14*, 391–402. Springer.

Balland, Pierre-Alexandre, Ron Boschma, Joan Crespo, and David L Rigby. 2018. "Smart Specialization Policy in the European Union: Relatedness, Knowledge Complexity and Regional Diversification." *Regional Studies*.

Balland, Pierre-Alexandre, Tom Broekel, Dario Diodato, Elisa Giuliani, Ricardo Hausmann, Neave O'Clery, and David Rigby. 2022. "The New Paradigm of Economic Complexity." *Research Policy* 51 (3): 104450.

Balland, Pierre-Alexandre, and David Rigby. 2017. "The Geography of Complex Knowledge." *Economic Geography* 93 (1): 1–23.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Boschma, Ron. 2017. "Relatedness as Driver of Regional Diversification: A Research Agenda." *Regional Studies* 51 (3): 351–64.

Broekel, Tom. 2019. "Using Structural Diversity to Measure the Complexity of Technologies." *PloS One* 14 (5): e0216856.

Chávez, Juan Carlos, Marco T Mosqueda, and Manuel Gómez-Zaldívar. 2017. "Economic Complexity and Regional Growth Performance: Evidence from the Mexican Economy." *Review of Regional Studies* 47 (2).

Cristelli, Matthieu, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. 2013. "Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products." *PloS One* 8 (8): e70726.

Griffiths, Thomas L, and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (suppl_1): 5228–35.

Guerzoni, Marco, Massimiliano Nuccio, and Federico Tamagni. 2024. "Pre-Entry Knowledge Base Complexity and Post-Entry Growth: Evidence from Italian Firms." *Industrial and Corporate Change* 33 (1): 126–51.

Hidalgo, César A. 2021. "Economic Complexity Theory and Applications." *Nature Reviews Physics* 3 (2): 92–113.

———. 2023. "The Policy Implications of Economic Complexity." *Research Policy* 52 (9): 104863.

Hidalgo, César A, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. 2007. "The Product Space Conditions the Development of Nations." *Science* 317 (5837): 482–87.

Ivanova, Inga, Nataliya Smorodinskaya, and Loet Leydesdorff. 2020. "On Measuring Complexity in a Post-Industrial Economy: The Ecosystem's Approach." *Quality & Quantity* 54: 197–212.

Ivanova, Inga, Øivind Strand, Duncan Kushnir, and Loet Leydesdorff. 2017. "Economic and Technological Complexity: A Model Study of Indicators of Knowledge-Based Innovation Systems." *Technological Forecasting and Social Change* 120: 77–89.

Kazemzadeh, Emad, José Alberto Fuinhas, Masoud Shirazi, Matheus Koengkan, and Nuno Silva. 2023. "Does Economic Complexity Increase Energy Intensity?" *Energy Efficiency* 16 (4): 29.

Koch, Philipp. 2021. "Economic Complexity and Growth: Can Value-Added Exports Better Explain the Link?" *Economics Letters* 198: 109682.

Nasios, Nikolaos, and Adrian G Bors. 2006. "Variational Learning for Gaussian Mixture Models." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36 (4): 849–62.

Neishabouri, Asana, and Michel C Desmarais. 2020. "Reliability of Perplexity to Find Number of Latent Topics." In *The Thirty-Third International Flairs Conference*.

Porteous, Ian, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation." In *Proceedings of the 14th ACM*

*SIGKDD International Conference on Knowledge Discovery and Data Mining*, 569–77.

Romero, João P, and Camila Gramkow. 2021. "Economic Complexity and Greenhouse Gas Emissions." *World Development* 139: 105317.

Sadeghi, Pegah, Hamid Shahrestani, Kambiz Hojabr Kiani, and Taghi Torabi. 2020. "Economic Complexity, Human Capital, and FDI Attraction: A Cross Country Analysis." *International Economics* 164: 168–82.

Tacchella, Andrea, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. 2012. "A New Metrics for Countries' Fitness and Products' Complexity." *Scientific Reports* 2 (1): 723.