Topic modeling and classification of scientific disciplines

Radim Hladik, Yann Renisio

Research in progress

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

OPEN ACCESS

Citation: Hladik, R., & Renisio, Y. (2022). Topic modeling and classification of scientific disciplines. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22223). https://doi.org/10.5281/zenodo.6957149

Collection: https://zenodo.org/communities/sti2022grx/

Topic modeling and classification of scientific disciplines[1]

Radim Hladík[*], Yann Renisio[**]

[*] *radim.hladik@fulbrightmail.org*
Institute of Philosophy, Czech Academy of Sciences, Jilská 1, Praha, 110 00 (Czech Republic)

[**] *yann.renisio@)sciencespo.fr*
Observatoire sociologique du changement, CNRS/Sciences-Po, 98 Rue de l'Université, Paris, 75007 (France)

Introduction

The choice of disciplinary classification systems has important consequences for the calculation of field-normalized scientometric indicators (Waltman & van Eck, 2019). However, despite their institutionalization through the organizational structure of research establishments and journals, disciplines are fluid, fractal categories (Abbott, 2001). In this paper, we evaluate the possibility of classifying Ph.D. theses into disciplines by using a bottom-up empirical approach based on topic modeling.

Data and methods

Our interest in the practical problem of the classification of scholarly communication into disciplinary categories stems from the experience of working with a dataset of 334810 Ph.D. theses submitted at French universities between 2006 and 2020. In this comprehensive dataset, the variable "discipline" does not rely on any controlled vocabulary or disciplinary nomenclature. Consequently, there are 23057 unique labels for the variable "discipline", of which 14538 appear only once. Such situation renders impossible any full-scale analysis of the data from the perspective of scientific disciplines.

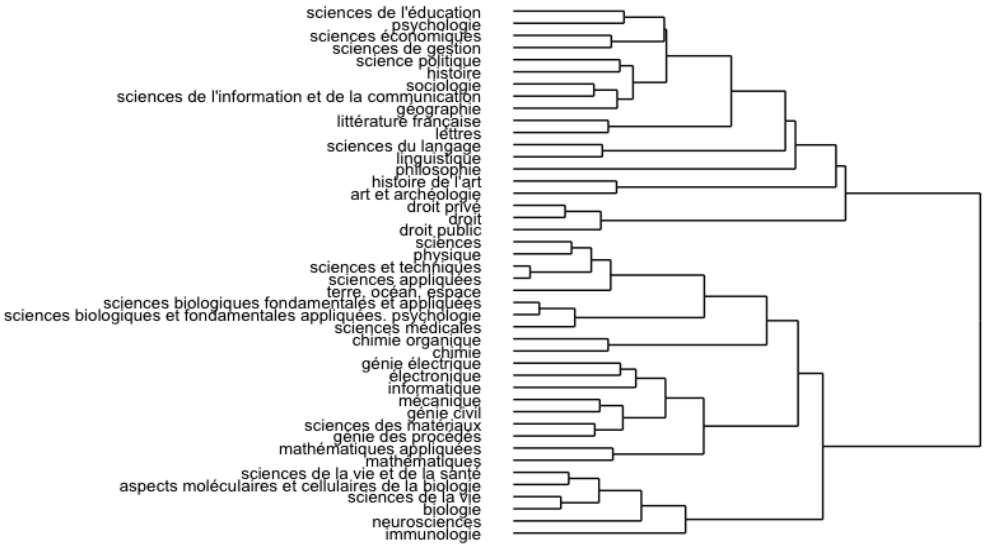
In the absence of citations data, cognitive and discursive dimensions of science can furnish insights into its intellectual organization (Leydesdroff, 1989; Foster, Rzhetsky, and Evans, 2015; Gerow et al., 2018; Dias et al., 2018). To account for the knowledge component of science, we rely on topic modeling, which ranks and clusters words based on their distribution in scientific publications. Topics are empirically constructed in an unsupervised manner and provide a middle ground, with a degree of semantic interpretability, between top-down, strictly delimited classification system and unstructured data. We present here the application of a methodological advance in topic modeling - the TopSBM algorithm (Gerlach, Peixoto, and Altmann, 2018), which employs hierarchical stochastic block modeling of communities in a bipartite network of words and documents to discover topics.

[1] This work was partially supported by the Czech Science Foundation project no. GJ20-01752Y, "Funded and Unfunded Research in the Czech Republic".

Our topic model is built atop of abstracts of 285311 of theses in French that include a title, keywords, and abstract. The texts were pre-processed by removing stopwords and non-alphabetical characters, creating compounds from frequent bi-grams and tri-grams, and lemmatization using UDPipe (Straka & Straková, 2017). After applying the TobSBM algorithm, we obtained a topic model with 7 levels of hierarchy. At the most nuanced level, the topic model contains 2043 topics to represent each document. For further steps, we employ the latter, most fine-grained solution.

Firstly, we explore the topic space of the theses by selecting discipline labels that appear at least 1000 times. There are 44 labels that satisfy this condition, and they encompass about half of the dataset, or 146099 documents. We calculate a mean topic vector for each discipline and hierarchically cluster them using the Ward's method. Figure 1 shows that the topic dimensions are highly structured. While it would be a challenge to ascertain the validity of the clustering solution, a qualitative overview indicates that similar disciplines are closely related via their topic features. The clusters also faithfully reproduce the great divide between natural and social sciences, as well as the differentiation among broad fields within them.

Figure 1: Clustering of most frequent disciplines in topic space meets qualitative expectations.



We next examine the ability of the topic features to predict labels of the most frequent disciplines to test if the topic model can be used beyond exploratory purposes. We eschew building a dedicated predictive model and, instead, proceed with a simple deterministic approach. First, we construct a reference dataset by sampling 10% of the theses from each discipline (n=14601) and average their topic vectors. We then use Kullback–Leibler divergence measure between two probability distributions to find, for each thesis in the remaining dataset (n= 131498) for testing purposes, its nearest disciplinary reference vector and assign it as the expected label.

Finally, to better understand the extent to which the topic model captures information about the disciplinary affiliation of the theses, we take advantage of the nested organizational structure of the French Conseil national des universités (CNS), which is consultative and decision-making body in the French system of tertiary education. CNS is organized in disciplinary sections which are themselves integrated into broad groups. The convoluted names of the CNS

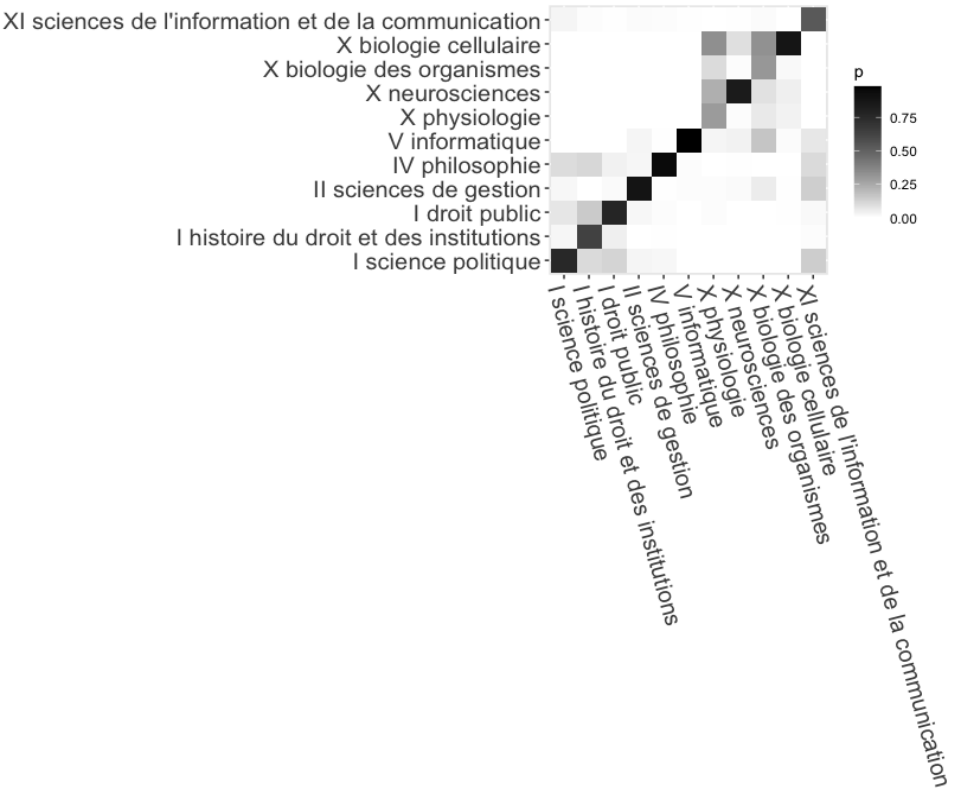
disciplinary sections do not map easily onto the labels in theses' dataset, but (with an additional requirement of having at least 10 observations in our reference dataset), we through exact matching, we could map 11 CNS sections to our data. For these 11 disciplines, there are 6 corresponding groups, 2 of which contain 3 and 4 disciplines respectively, while the remaining 4 disciplines are each in their own group. Next, we again construct a reference dataset (n= 2920) for the subset of CNS-matched labels and, with its help, attempt to determine the discipline of the theses in the remainder of the data (n=26066).

Results

Our first experiment with the most frequent disciplines yielded a classification accuracy of 47%. While this is a very satisfactory result for a problem with 44 classes and a solution without a black-boxed model, it may not suffice for practical application in imputing disciplinary labels. More evidence of the usefulness of the topic modelling is required.

The results for our second experiment appear in Figure 2, where disciplines are sorted in the arbitrary order of the CNS's system of disciplinary groups (designated here by Roman numerals). Where only 1 discipline per CNS group could be found in the data, misclassifications are rare. Most disciplines get misclassified within their respective encompassing groups, which suggests that even the wrongly predicted labels remain, in fact, substantively valid and close to the ground truth. Overlaps appear between such arguably related disciplines as "public law" and "history of law and institutions" within group I, or "cellular biology" and "physiology" within group X. The topic features sometimes mistake "information and communication sciences" for other humanities or social-scientific disciplines, but this discipline itself belongs to a group designated by the CNU as "multidisciplinary".

Figure 2: Misclassified disciplines are mostly contained within their official groups.



For 11 disciplinary classes, our approach achieved 86% classification accuracy, and for the 6 disciplinary groups, the metric is 91%. Only 9% of the theses were therefore attributed to disciplines outside of their actual disciplinary group. Instead of being simple failures, the instances of incorrect labels assigned from the topic space may therefore reveal tensions between the actual intellectual content of the work and the choices and constraints that students face when they formally position their work in the system of academic disciplines.

Discussion and Conclusion

The outcomes of both experiments suggest that topics derived from purely textual data implicitly capture information about disciplines. This quality of topic modelling can be of great benefit when dealing with datasets where disciplinary information is unavailable or unreliable and where citation records are absent (as it remains the case especially in the Humanities). Even if topics cannot fully reconstruct originally assigned disciplines, they still provide reliable pointers about the broader field to which a document belongs. Another advantage is that after training a topic model, no further models need to be built because the deterministic divergence measure offers an adequate solution. However, a wider range of metrics and experiments with other datasets will be needed to fully assess the performance of topics as predictors of disciplines.

The results of this preliminary analysis allow for two possible conjunctures. A conservative perspective posits that topic models provide a representation of scholarly documents that researchers can use either to deterministically predict disciplinary labels or to develop clustering solutions to plausibly mirror a system of scientific disciplines. A more radical implication is that proportional assignment of scholarly work to empirically constructed topics provides a viable alternative to strict disciplinary classifications of various provenience. Topic models seem to be well-suited for capturing the inevitable fluidity and interdisciplinarity of scientific knowledge production. Methodologically, we propose that nested systems of classification offer an expedient framework for the evaluation of automated predictions of labels with fuzzy boundaries, such as scientific disciplines.

References

Abbott, A. (2001). *Chaos of disciplines*. University of Chicago Press.

Dias, L., Gerlach, M., Scharloth, J., & Altmann, E. G. (2018). Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, *5*(1), 171545.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, *4*(7), eaaq1360.

Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, *115*(13), 3308–3313.

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, *18*(4), 209–223.

Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Waltman, L., & van Eck, N. J. (2019). Field Normalization of Scientometric Indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 281–300). Springer International Publishing.