Digitalization and the Organization of Work in the Digital Government Research Library v17.5: A Structural Topic Model

Andres Aguilera Castillo

10 May 2022

Abstract

Digital Government is a growing and vibrant multidisciplinary field of research, the fast increase in research output has challenged researchers to explore and use novel computational ways and methods to perform evidence synthesis on the extant literature and be able to map a scientific discipline, explore the thematic evolution over time and identify potential avenues for further research. The exploration of the linkages between digitalization and the organization of work remains relatively unexplored in a public sector context. Topic modeling has emerged as a powerful computational technique that contributes to the examination of large amounts of text data. This chapter aims to perform a 'smart literature review' on a subset of the Digital Government Reference Library (DGRL) version 17.5, by using text as data approach and a topic modeling technique known as Latent Dirichlet Allocation (LDA). To our best knowledge, this is the first attempt to use unsupervised machine learning techniques with this data set. This effort may contribute to creating a map of the field, identify the evolving themes in the literature and help to identify promising areas of research.

Introduction

Recent trends in global scientific output demonstrate a rapid and sustained increase in the production of vast amounts of unstructured data in the form of digitized text. This bounty in research content is challenging researchers to explore and pursue novel methodological approaches and techniques to examine massive volumes of scientific publications. Essentially, the ever-growing amounts of bibliographic information available in almost any field of research exceed human capacity making it necessary to explore computational-assisted approaches for research synthesis. Studies in the history of science have identified a relatively sustained growth pattern in scientific publications over time, this exponential growth rate means a doubling in scientific output every 17 years approximately (Bornmann, Haunschild, and Mutz 2021). This level of growth might be attributed to the increased resources dedicated to the global scientific endeavor and consequently the communication of science via publications. However, it may also be due to what has been dubbed "salami sliced publishing" or the multiple publications of a single research study (Bornmann and Daniel 2007; Bornmann and Mutz 2015).

Digital Government Research (DGR) as a multidisciplinary research field is experiencing rapid growth in its research output. Contributions to this research domain come from established disciplines such as information science, computer science, organization science, sociology, public administration, and political science (Scholl 2021). The diversity in scope and methods of these disciplines converge in the field of Digital Government enriching it, but at the same time, raising questions as regards the lack of native theoretical developments, thus relying upon frameworks, theories, and conceptualizations from related disciplines (Bannister and Connolly 2015).

The advent of computerization and digitalization has had broad impacts in most aspects of contemporary life, including scientific research. Digitalization has influenced how research is designed and conducted, allowing for the creation and increased availability of ever-growing data sets that require powerful computational methods and enhanced tools to handle abundant information (Meyer and Schroeder 2015). Case in point is unsupervised machine learning techniques for text analysis, this research technology can be used in a wide range of disciplines to examine databases, repositories and corpora thus expanding the methodological repertoire of scholars and opening an opportunity to explore large troves of data.

Research synthesis is part of the literature review process in which the extant scientific knowledge in each academic field is examined to help scholars understand the conceptual structure, themes, and debates to identify trends in the literature and potential areas for further research. This crucial task is labor-intensive, time-consuming, and restricted to a limited number of documents if conducted by traditional "manual" methods (Antons et al. 2020a; Asmussen and Møller 2019). Still, computer-assisted text analysis does not substitute human intervention, instead it "augments our reading ability" (Grimmer, Roberts, and Stewart 2022a), human judgement is deemed necessary for the evaluation and validation of the outcome of these models (Barberá et al. 2021).

This chapter explores the contents of The Digital Government Research Library in its version 17.5 (Scholl, 2021a), using a text as data approach and an unsupervised machine learning technique known as Latent Dirichlet Allocation (LDA). According to the DGRL website, version 17.5 of the data set contains 16531 references related to the Digital Government Research (DGR) domain, the most prevalent types of documents are conference papers (33.2%) and journal articles (50%).

Table 1: Table 1. Contents of DGRL v17.5 without pre-processing

Document Type	Number of Documents
journalArticle	8278
conferencePaper	5492
bookSection	2083
book	636
report	33
thesis	3
magazineArticle	2
manuscript	1
webpage	1

Previous explorations of this reference library using bibliometric and scientometric approaches have revealed the thematic evolution (AlcaideMuñoz et al. 2017), and identified the most influential journals, conferences and leading scholars in the field (Scholl, 2021b). This data set has been used as well for conducting a systematic review on the impacts of e-Government using a public value perspective (MacLean and Titah 2021). To our best knowledge this is the first attempt to run a topic model for a corpus in the field of Digital Government Research.

Scholars exploring labor-saving technologies have applied similar techniques expanding the methodological repertoire available and inspiring a similar pursuit for the exploration of the impact of technological change in a public sector context (Montobbio et al. 2022).

RQ

Process

Literature Review // Conceptual framework

Quantitative research synthesis techniques like bibliometrics and computer-assisted text mining allow the analysis of a larger quantity of documents and may contribute to advancing the "research fronts" in interdisciplinary fields such as Digital Government (Tanskanen et al. 2017). Computational tools and techniques developed originally in the computer science field have been repurposed in diverse disciplines but also have enabled social scientists to exploit Natural Language Processing (NLP) applications for classification tasks of large scientific corpora. Topic modeling techniques, a subset of machine learning and NLP allow for the automatic classification of vast amounts of text and have been used for the analysis of bibliographic content in diverse fields of research and academic disciplines including statistics (De Battisti, Ferrara, and Salini 2015), economics (Ambrosino et al. 2018), cliometrics (Wehrheim 2019), innovation research (Antons et al. 2020b; Antons and Breidbach 2017), and management (Hannigan et al. 2019).

On occasions the scope of analysis can be very large, (Ambrosino et al. 2018) studied the evolution in the thematic structure of the economics discipline by applying LDA to the full texts of articles published in 188 journals in the JSTOR database from 1845 to 2013 (n= 250846). Other implementations of these techniques (Antons, Kleer, and Salge 2016), have explored the full text corpus of a single top ranking journal in innovation research over a three decade span (n=1008), alternative uses of this technique have considered the titles of dissertations in economics and chemistry in East and West Germany before and after the German reunification (Rehs 2020).

As advised by (Barberá et al. 2021), there are "consequential decisions" in the methodological choices of automated text classification and the fact that human validation is a key component of text as data methods. The selection of a corpus in itself is deemed a crucial decision that can be prone to four types of bias: resource bias, incentive bias, medium bias and retrieval bias, these selection biases are well acknowledged in the text as data literature (Grimmer, Roberts, and Stewart 2022a). The inclusion criteria of the DGRL are: to have passed academic peer review, to be published in an academic journal, to be published in English language (Scholl 2021). The criteria described above may help to mitigate the potential selection biases, the DGRL v17.5 may probably ignore or leave out important research out of its scope, however, echoing Barberà et al, all decisions concerning text as data methodologies are "consequential" and our aim is to make our workflow reproducible by documenting the choices in the scripts associated with this document.

The corpus used for this analysis is a subset of the journal articles in the version 17.5 of the Digital Government Reference Library. As argued by Grimmer, Roberts, and Stewart (2022b), texts are "expensive to produce, gather and collate", the contents of previous versions of this data set have been used as primary or secondary source of data exploring the Digital Government field. The approach of this chapter is not as ambitious as Ambrosino *et. al.*, nor focalized on a single journal publication as Antons *et. al.*, instead the aim is to analyze the abstracts of 6682 journal articles in the version 17.5 of the DGRL via the application of a structural topic model.

Evidence synthesis –>

Collect names (Buffat et al)

STM

reproducibility

human interpretable

Insert list of top journals.

Methods and Data

The Digital Government Research Library version 17.5 is a large curated repository of publications contributing to the field of Digital Government Research (DGR), it contains more than 16500 references among its records. The Library can be downloaded from the website DGRL. The download package contains three types of bibliographic files BibTeX, RIS, and ENL (EndNote). In its raw and unprocessed form, the data has a large proportion of missing values, mostly clustered in metadata not considered relevant for the analysis. By exploring the BiBTeX, RIS and ENL files, we noticed that the data sets had a large amount of missing data and that some information was available in a file type but not other. For this exercise, the following variables have been deemed of interest for the analysis: type of reference (conference paper or journal article), year of publication, author, document title, publication title and the presence of an abstract. Text is a type of unstructured data that requires meticulous processing before using it. For replicability purposes, the script for data wrangling, cleaning and overall processing of the topic model is available in the scripts section of the GitHub repository for this project.

After the initial data wrangling, the relevant data for 6682 journal articles or approximately 80.7% of the total number of articles in the DGRL v17.5 is further processed to create a corpus, the initial step towards a topic model. A visualization in the publication trend demonstrate an incipient increase in number of journal articles after year 2000 and a steep increase in the beginning of the 2010 decade to present.

The workflow for topic modeling includes text pre-processing, meaning further data cleaning and data transformation. This means that before creating a corpus object with the available information from the DGRL v17.5, the text should be prepared before running the initial explorations. Our main unit of analysis is the abstract of the journal articles contained in the processed data set of the DGRL v17.5. Text as a type of unstructured data can be structured for processing using the bag of words approach or the splitting of abstracts into separate word units or terms and every occurrence of a term is defined as a token. The creation of a bag of words is known as tokenizing. The bag of words approach deliberately ignores the syntax or structure of the text, additional treatment of text include the elimination of punctuation, transform each word to lowercase and in some cases stemming which is a way to reduce a word to its stem or root.

Stemming algorithms: Porter / Snowball / Lancaster

Stop words are context specific

Dictionary based Quantitative Text Analysis Based on the work of Montobbio et al. 2022

Truncated words

Theory based Watanabe.

Results

Discussion

Conclusions

References

AlcaideMuñoz, Laura, Manuel Pedro RodríguezBolívar, Manuel Jesús Cobo, and Enrique HerreraViedma. 2017. "Analysing the Scientific Evolution of e-Government Using a Science Mapping Approach." *Government Information Quarterly* 34 (3): 545–55. https://doi.org/10.1016/j.giq.2017.05.002.

Ambrosino, Angela, Mario Cedrini, John B. Davis, Stefano Fiori, Marco Guerzoni, and Massimiliano Nuccio. 2018. "What Topic Modeling Could Reveal about the Evolution of Economics." *Journal of Economic Methodology* 25 (4): 329–48. https://doi.org/10.1080/1350178X.2018.1529215.

Antons, David, and Christoph F. Breidbach. 2017. "Big Data, Big Insights? Advancing Service Innovation and Design with Machine Learning:" *Journal of Service Research* 21 (1): 17–39. https://doi.org/10.1177/1094670517738373.

Antons, David, Eduard Grünwald, Patrick Cichy, and Torsten Oliver Salge. 2020a. "The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities." *R&D Management* 50 (3): 329–51. https://doi.org/10.1111/RADM.12408.

———. 2020b. "The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities." *R&D Management* 50 (3): 329–51. https://doi.org/10.1111/RADM.12408.

Antons, David, Robin Kleer, and Torsten Oliver Salge. 2016. "Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories." *Journal of Product Innovation Management* 33 (6): 726–49. https://doi.org/10.1111/jpim.12300.

Asmussen, Claus Boye, and Charles Møller. 2019. "Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review." *Journal of Big Data* 6 (1): 1–18. https://doi.org/10.1186/S40537-019-0255-7/TABLES/6.

Bannister, Frank, and Regina Connolly. 2015. “The Great Theory Hunt: Does e-Government Really Have a Problem?” *Government Information Quarterly* 32 (1): 1–11. https://doi.org/10.1016/J.GIQ.2014.10.003.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. “Automated Text Classification of News Articles: A Practical Guide.” *Political Analysis* 29 (1): 19–42. https://doi.org/10.1017/PAN.2020.8.

Bornmann, Lutz, and Hans-Dieter Daniel. 2007. “Multiple Publication on a Single Research Study: Does It Pay? The Influence of Number of Research Articles on Total Citation Counts in Biomedicine.” *Journal of the American Society for Information Science and Technology* 58 (8): 1100–1107. https://doi.org/10.1002/ASI.20531.

Bornmann, Lutz, Robin Haunschild, and Rüdiger Mutz. 2021. “Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases.” *Humanities and Social Sciences Communications 2021 8:1* 8 (1): 1–15. https://doi.org/10.1057/s41599-021-00903-w.

Bornmann, Lutz, and Rüdiger Mutz. 2015. “Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References.” *Journal of the Association for Information Science and Technology* 66 (11): 2215–22. https://doi.org/10.1002/asi.23329.

De Battisti, Francesca, Alfio Ferrara, and Silvia Salini. 2015. “A Decade of Research in Statistics: A Topic Model Approach.” *Scientometrics 2015 103:2* 103 (2): 413–33. https://doi.org/10.1007/S11192-015-1554-1.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022a. *Text as data: a new framework for machine learning and the social sciences.* Princeton, New Jersey Oxford: Princeton University Press.

———. 2022b. *Text as data: a new framework for machine learning and the social sciences.* Princeton, New Jersey Oxford: Princeton University Press.

Hannigan, Timothy R., Richard F. J. Haan, Keyvan Vakili, Hovig Tchalian, Vern L. Glaser, Milo Shaoqing Wang, Sarah Kaplan, and P. Devereaux Jennings. 2019. “Topic Modeling in Management Research: Rendering New Theory from Textual Data.” *Academy of Management Annals* 13 (2): 586–632. https://doi.org/10.5465/ANNALS.2017.0099.

MacLean, Don, and Ryad Titah. 2021. “A Systematic Literature Review of Empirical Research on the Impacts of e-Government: A Public Value Perspective.” *Public Administration Review*, August. https://doi.org/10.1111/PUAR.13413.

Meyer, Eric T., and Ralph Schroeder. 2015. *Knowledge Machines: Digital Transformations of the Sciences and Humanities.* Infrastructures. Cambridge, Massachusetts: The MIT Press.

Montobbio, Fabio, Jacopo Staccioli, Maria Enrica Virgillito, and Marco Vivarelli. 2022. “Robots and the Origin of Their Labour-Saving Impact.” *Technological Forecasting and Social Change* 174 (January): 121122. https://doi.org/10.1016/J.TECHFORE.2021.121122.

Rehs, Andreas. 2020. “A Structural Topic Model Approach to Scientific Reorientation of Economics and Chemistry After German Reunification.” *Scientometrics* 125 (2): 1229–51. https://doi.org/10.1007/S11192-020-03640-0/TABLES/4.

Scholl, Hans J. 2021. “The Digital Government Reference Library (DGRL) and Its Potential Formative Impact on Digital Government Research (DGR).” *Government Information Quarterly*, July, 101613. https://doi.org/10.1016/J.GIQ.2021.101613.

Tanskanen, Kari, Tuomas Ahola, Anna Aminoff, Johanna Bragge, Riikka Kaipia, and Katri Kauppi. 2017. “Towards Evidence-Based Management of External Resources: Developing Design Propositions and Future Research Avenues Through Research Synthesis.” *Research Policy* 46 (6): 1087–1105. https://doi.org/10.1016/j.respol.2017.04.002.

Wehrheim, Lino. 2019. “Economic History Goes Digital: Topic Modeling the Journal of Economic History.” *Cliometrica* 13 (1): 83–125. https://doi.org/10.1007/S11698-018-0171-7/TABLES/4.