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

Andres Aguilera Castillo

12 April 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. 2020; Asmussen and Møller 2019). Still, computer-assisted text analysis does not substitute human intervention, instead it "augments our reading ability" (Grimmer, Roberts, and Stewart 2022), 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). In its raw form, version 17.5 of the dataset contains 16531 references related to the Digital Government Research (DGR) domain, the most prevalent categories of references are conference papers (33%) and journal articles (50%). Previous explorations of this reference library using bibliometric and scientometric approaches have revealed the thematic evolution (Alcaide–Muñoz et al., 2017), and identified the most influential journals, conferences and leading scholars in the field (Scholl, 2021b). 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 (Arduini and Zanfei 2014; Montobbio et al. 2022).

RQ

Process

```
summary(DGRLv17_5_zotero)
```

```
##      Key              Item Type          Publication Year    Author
##  Length:16531      Length:16531       Min.   :1981       Length:16531
##  Class :character  Class :character   1st Qu.:2008       Class :character
##  Mode  :character  Mode  :character   Median :2013       Mode  :character
##                                       Mean   :2013
##                                       3rd Qu.:2018
##                                       Max.   :2022
##                                       NA's   :4165
##     Title          Publication Title       ISBN              ISSN
##  Length:16531      Length:16531       Length:16531       Length:16531
##  Class :character  Class :character   Class :character   Class :character
##  Mode  :character  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      DOI               Url             Abstract Note        Date
##  Length:16531      Length:16531       Length:16531       Length:16531
##  Class :character  Class :character   Class :character   Class :character
##  Mode  :character  Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##    Date Added                 Date Modified
## Min.   :2022-02-17 12:47:09   Min.   :2022-02-17 12:47:09
## 1st Qu.:2022-02-17 12:51:19   1st Qu.:2022-02-17 12:51:19
## Median :2022-02-17 12:55:04   Median :2022-02-17 12:55:04
## Mean   :2022-02-17 12:55:16   Mean   :2022-02-17 12:55:16
## 3rd Qu.:2022-02-17 12:59:11   3rd Qu.:2022-02-17 12:59:11
## Max.   :2022-02-17 13:03:57   Max.   :2022-02-17 13:03:57
##
##    Access Date                    Pages            Num Pages
## Min.   :0006-01-01 00:00:00   Length:16531      Length:16531
## 1st Qu.:2015-11-05 00:00:00   Class :character   Class :character
## Median :2017-11-07 00:00:00   Mode  :character   Mode  :character
## Mean   :2016-08-26 08:09:30
## 3rd Qu.:2019-05-07 18:41:30
## Max.   :2020-12-29 18:46:28
## NA's   :13705
##    Issue              Volume          Number Of Volumes Journal Abbreviation
## Length:16531       Length:16531       Min.   :1.000      Length:16531
## Class :character   Class :character   1st Qu.:2.000      Class :character
## Mode  :character   Mode  :character   Median :3.000      Mode  :character
##                                       Mean   :2.732
##                                       3rd Qu.:3.000
##                                       Max.   :3.000
##                                       NA's   :16337
## Short Title          Series          Series Number     Series Text
## Length:16531       Length:16531       Length:16531      Mode:logical
## Class :character   Class :character   Class :character  NA's:16531
## Mode  :character   Mode  :character   Mode  :character
##
##
##
##
## Series Title     Publisher            Place            Language
## Mode:logical    Length:16531        Length:16531      Length:16531
## NA's:16531      Class :character    Class :character   Class :character
##                 Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##  Rights           Type              Archive           Archive Location
## Mode:logical    Mode:logical     Length:16531       Length:16531
## NA's:16531      NA's:16531       Class :character   Class :character
##                                  Mode  :character   Mode  :character
##
##
##
##
## Library Catalog    Call Number          Extra              Notes
## Length:16531       Length:16531        Length:16531       Length:16531
## Class :character   Class :character    Class :character   Class :character
```

```
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
## File Attachments Link Attachments Manual Tags        Automatic Tags
## Mode:logical     Mode:logical     Length:16531       Mode:logical
## NA's:16531       NA's:16531       Class :character   NA's:16531
##                                   Mode  :character
##
##
##
##
##    Editor           Series Editor      Translator      Contributor
## Length:16531       Length:16531       Mode:logical    Mode:logical
## Class :character   Class :character   NA's:16531      NA's:16531
## Mode  :character   Mode  :character
##
##
##
##
## Attorney Agent Book Author    Cast Member    Commenter      Composer
## Mode:logical    Mode:logical    Mode:logical    Mode:logical    Mode:logical
## NA's:16531      NA's:16531      NA's:16531      NA's:16531      NA's:16531
##
##
##
##
## Cosponsor       Counsel         Interviewer     Producer        Recipient
## Mode:logical    Mode:logical    Mode:logical    Mode:logical    Mode:logical
## NA's:16531      NA's:16531      NA's:16531      NA's:16531      NA's:16531
##
##
##
##
## Reviewed Author Scriptwriter    Words By        Guest           Number
## Mode:logical    Mode:logical    Mode:logical    Mode:logical    Mode:logical
## NA's:16531      NA's:16531      NA's:16531      NA's:16531      NA's:16531
##
##
##
##
##    Edition          Running Time    Scale           Medium          Artwork Size
## Length:16531       Mode:logical    Mode:logical    Mode:logical    Mode:logical
## Class :character   NA's:16531      NA's:16531      NA's:16531      NA's:16531
## Mode  :character
##
##
##
##  Filing Date     Application Number Assignee        Issuing Authority
```

```
##    Mode:logical    Mode:logical       Mode:logical    Mode:logical
##    NA's:16531      NA's:16531         NA's:16531      NA's:16531
##
##
##
##
##
##    Country         Meeting Name   Conference Name     Court          References
##    Mode:logical    Mode:logical   Length:16531        Mode:logical   Mode:logical
##    NA's:16531      NA's:16531     Class :character    NA's:16531     NA's:16531
##                                   Mode  :character
##
##
##
##
##    Reporter        Legal Status   Priority Numbers Programming Language
##    Mode:logical    Mode:logical   Mode:logical       Mode:logical
##    NA's:16531      NA's:16531     NA's:16531         NA's:16531
##
##
##
##
##
##    Version         System         Code            Code Number    Section
##    Mode:logical    Mode:logical   Mode:logical    Mode:logical   Mode:logical
##    NA's:16531      NA's:16531     NA's:16531      NA's:16531     NA's:16531
##
##
##
##
##
##    Session         Committee      History         Legislative Body
##    Mode:logical    Mode:logical   Mode:logical    Mode:logical
##    NA's:16531      NA's:16531     NA's:16531      NA's:16531
##
##
##
##
##
```

Literature Review // Conceptual framework

You can also embed plots, for example:

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), in its raw form this version of the library has 16531 references, the lion's share being journal articles n=8278 and conference papers n=5493.

```
by_type <- DGRLv17_5_zotero %>% count(`Item Type`) %>% arrange(desc(n))
by_type
```

```
## # A tibble: 9 x 2
##   `Item Type`        n
##   <chr>          <int>
## 1 journalArticle  8278
## 2 conferencePaper 5493
## 3 bookSection     2084
## 4 book             636
## 5 report            33
## 6 thesis             3
## 7 magazineArticle    2
## 8 manuscript         1
## 9 webpage            1
```

```
vis_miss(DGRLv17_5_zotero, warn_large_data = FALSE)
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```
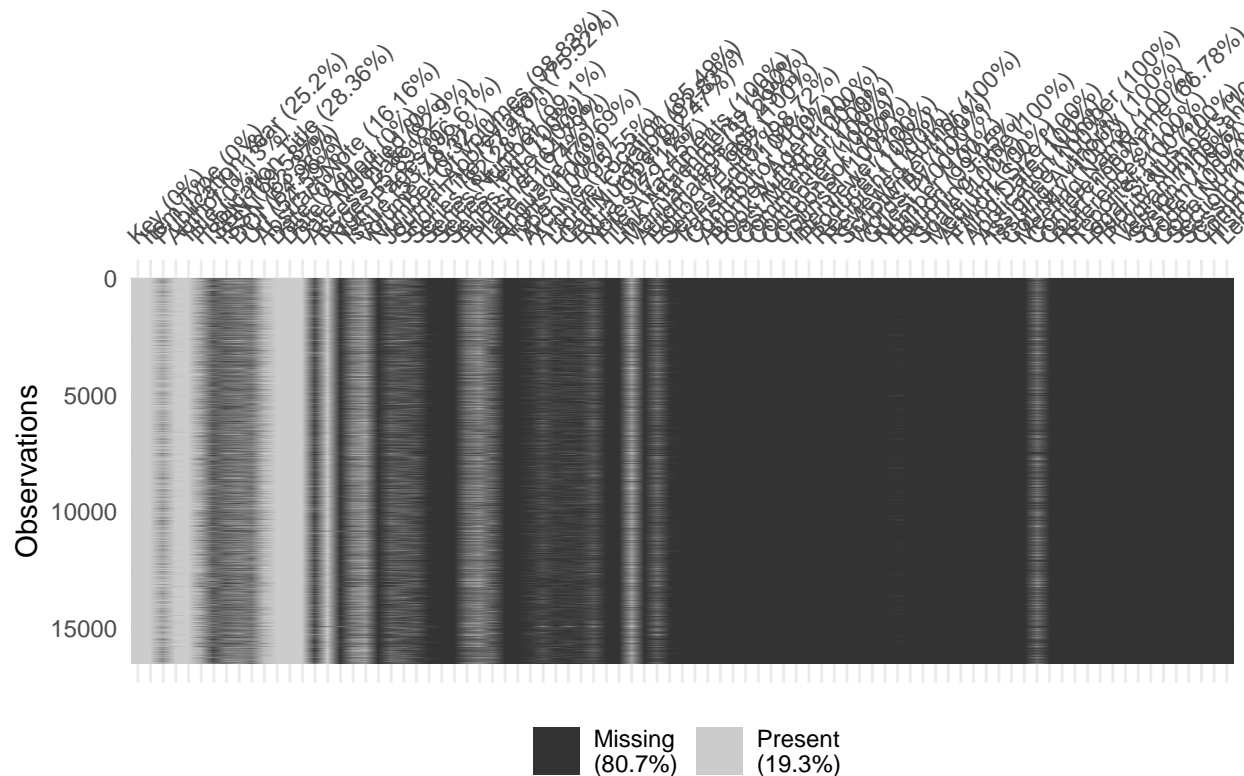
```
miss_all_db <- pct_miss(DGRLv17_5_zotero)
```

The data set has a large proportion of missing values 80.7470048 but not all columns are relevant to this exploration.

By visually exploring a reduced version of the dataset, it is evident that a large amount of relevant information is missing.

The workflow for reproducibility includes text pre-processing meaning data cleaning and data transformation. Some variables of interest are missing and there must be some data preparation before running creating a corpus.

Results

Discussion

Conclusions

References

Antons, David, Eduard Grünwald, Patrick Cichy, and Torsten Oliver Salge. 2020. "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.

Arduini, Davide, and Antonello Zanfei. 2014. "An Overview of Scholarly Research on Public e-Services? A Meta-Analysis of the Literature." *Telecommunications Policy* 38 (5-6): 476–95. https://doi.org/10.1016/j.telpol.2013.10.007.

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.

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

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.

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.