The Journal of Open Source Software

text2sdg: An open-source solution to monitoring sustainable development goals from text

Dominik S. Meier[1], **Rui Mata**[1, 2], **and Dirk U. Wulff**[1, 2]

1 University of Basel **2** Max Planck Institute for Human Development

DOI: 10.21105/joss.03988

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 10 December 2021
Published: 11 December 2021

Summary

Monitoring progress on the United Nations Sustainable Development Goals (SDGs) is important for both academic and non-academic organizations. Existing approaches to monitoring SDGs have focused on specific data types, namely, publications listed in proprietary research databases. We present the `text2sdg` R package, a user-friendly, open-source package that detects SDGs in any kind of text data using several scientifically-developed query systems. The `text2sdg` package facilitates the monitoring of SDGs for a wide array of text sources and provides a much-needed basis for validating and improving extant methods to detect SDGs in text.

Statement of need

The United Nations' Sustainable Development Goals (SDGs) have become an important guideline for both governmental and non-governmental organizations to monitor and plan their contributions to social, economic, and environmental transformations. As the latest UN report (UN, 2021) attests, progress is still needed and the availability of high-quality data will be critical to identify areas requiring most attention moving forward. One promising way to monitor progress on SDGs is to screen the increasing amount of digitally available text using automatized, natural language processing methods. This approach has taken hold, for example, in scientometric efforts that monitor the SDGs in academic publications (e.g., Jayabalasingham et al., 2021). These efforts have so far been spearheaded by for-profit organizations with methodologies that are only partly publicly available and cannot be easily applied beyond academic publishing databases. In what follows, we describe some of these approaches and discuss their shortcomings before introducing our open-source solution, the `text2sdg` R package, which is designed to help monitor work on the sustainable development goals from any text source.

There are currently five leading approaches to monitoring SDGs from text. The most influential of these was developed by the Elsevier SDG Research Mapping Initiative and uses Lucene-like queries to map several features available from a proprietary database of scientific publications (i.e., Scopus; https://www.scopus.com), including most importantly the abstracts of the publications, to SDGs (Jayabalasingham et al., 2021). The Elsevier system has been found to detect millions of SDG-related publications (Agnew et al., 2020) and is used by the Times Higher Education Impact Rankings to rank over 1000 universities worldwide according to their SDG-related research output. Elsevier recently partnered with Aurora, a network of universities that had independently developed a query system to detect SDGs in publications (Vanderfeesten et al., 2020). Other extant query systems include those of OSDG (Bautista, 2019), Siris (Duran-Silva et al., 2019), and the Sustainable Development Solutions Network (Sustainable Development Solutions Network, 2021). Despite the effort put into developing

these systems, they are not without shortcomings. First, they are not directly applicable to text sources other than academic citation databases (e.g., Scopus). Second, the systems lack user-friendly and transparent ways to communicate which features are matched to which documents or how they compare between a choice of query systems.

We alleviate these shortcomings by providing an open-source solution, `text2sdg`, that lets users detect SDGs in any kind of text using any of the above-mentioned systems or, even, customized, user-made query systems. The package provides a common framework to implement the different extant approaches and makes it easy to quantitatively compare and visualize their results.

Features

We showcase the potential of `text2sdg` with an analysis of a publicly available database of research projects funded by the Swiss National Science Foundation (https://p3.snf.ch). To do this, we first applied the packages' main function, `detect_sdg()`, to 26,811 English abstracts written by the research project authors, while setting the `system` argument to `c("Aurora", "SIRIS", "Elsevier", "SDSN", "OSDG")` in order to recruit all five query systems. We then used the packages' `plot_sdg()` function to visualize the frequency of SDG hits and the `crosstab_sdg()` function to analyze the correspondence between query systems and SDGs.

Figure 1 shows the results. We highlight three main findings. First and foremost, the results show it is possible to systematically map SDGs to text from sources other than citation databases. Second, as can be seen in panels B-D, the results suggest important and sizable differences between query systems in both the number of hits and the profiles of most researched SDGs, suggesting it is important to question the results from any single approach. Third, the results suggest that the SDGs vary considerably in their overlap (panel E), emphasizing the promise and challenges of tackling different SDGs simultaneously.

The Journal of Open Source Software

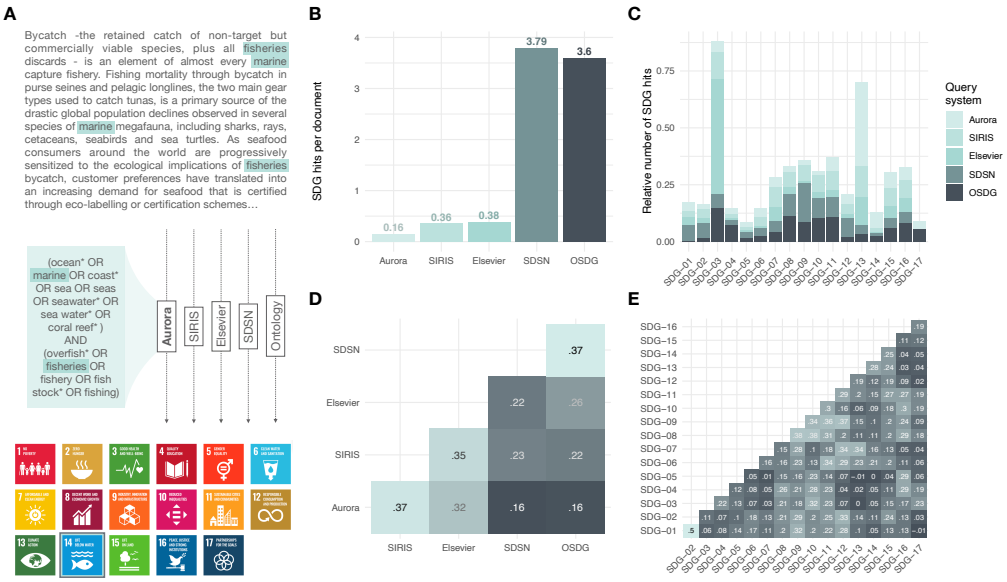

Figure 1: Analysis of 26,811 research projects funded by the Swiss National Science Foundation using the five query systems currently available in `text2sdg`. Panel A illustrates the identification of SDGs based on a query search: The example shows an excerpt of one abstract with some terms highlighted that match a query of the Aurora query system indicating SDG-14 (i.e., conserve and sustainably use the oceans, seas and marine resources for sustainable development). Panel B shows the number of hits per document for the different systems made available in `text2sdg`, which reveals striking differences in numbers of hits. Panel C shows the relative number of hits per SDG cumulatively across systems. Panel D shows the correlations between query systems, which reveal overall small to medium levels of correspondence between them. Panel E shows the correlation between detected SDGs over all query systems.

To facilitate further development of SDG query systems, `text2sdg` adidtionally includes the `detect_any` function, which permits users to run self-specified queries. We hope this function can be instrumental in future efforts to advance work that validates existing and future query systems and, therefore, can contribute to advancing the sustainable development goals.

Learn more about the background and functionality of the package at https://dwulff.github.io/text2sdg. See the package vignette (https://dwulff.github.io/text2sdg/articles/text2sdg.html), which includes a reproducible analysis pipeline involving all functions and the package's `projects` dataset consisting of a random subset of research project abstracts.

References

Agnew, K., Francescon, D., Martin, R., Rhannam, M., & Schemm, Y. (2020). *The Power of Data to Advance the SDGs. Mapping research for the Sustainable Development Goals*. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0004/1058179/Elsevier-SDG-Report-2020.pdf

Bautista, N. (2019). *SDG ontology* [Data set]. https://doi.org/10.6084/m9.figshare.11106113.v1

Duran-Silva, N., Fuster, E., Massucci, F. A., & Quinquillà, A. (2019). *A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals* (Version 1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3567769

Jayabalasingham, B., Boverhof, R., Agnew, K., & Klein, L. (2021). *Identifying research supporting the United Nations sustainable development goals* [Data set]. https://doi.org/10.17632/87txkw7khs.1

Sustainable Development Solutions Network. (2021). *Compiled list of SDG keywords* [Data set]. https://ap-unsdsn.org/regional-initiatives/universities-sdgs/

UN. (2021). *The Sustainable Development Goals Report 2021*. United Nations. https://doi.org/10.18356/9789210056083

Vanderfeesten, M., Otten, R., & Spielberg, E. (2020). *Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0* (Version 5.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3817445