Package 'text2sdg'

March 17, 2023

Type Package

Version 1.1.1

Date 2023-3-16

Title Detecting UN Sustainable Development Goals in Text

Description
The United Nations' Sustainable Development Goals (SDGs) have become an important guideline for organisations to monitor and plan their contributions to social, economic, and environmental transformations. The 'text2sdg' package is an open-source analysis package that identifies SDGs in text using scientifically developed query systems, opening up the opportunity to monitor any type of text-based data, such as scientific output or corporate publications. For more information regarding the methodology see Meier, Mata & Wulff (2022) <arXiv:2110.05856>.

Maintainer Dominik S. Meier <dominikmeier@outlook.com>

URL https://github.com/dwulff/text2sdg

BugReports https://github.com/dwulff/text2sdg/issues

License GPL-3

Encoding UTF-8

Depends R (>= 3.5.0)

Imports magrittr, dplyr, corpustools (>= 0.4.2), tidyr, tibble, stringr, ggplot2, lifecycle, ranger, text2sdgData (>= 0.1.1)

Suggests testthat (>= 3.0.0), knitr, rmarkdown

LazyData TRUE

LazyDataCompression bzip2

RoxygenNote 7.2.1

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author Dirk U. Wulff [aut] (<https://orcid.org/0000-0002-4008-8022>), Dominik S. Meier [aut, cre] (<https://orcid.org/0000-0002-3999-1388>), Rui Mata [ctb] (<https://orcid.org/0000-0002-1679-906X>)

Repository CRAN

Date/Publication 2023-03-17 20:50:02 UTC

R topics documented:

`auckland_queries` *SDG queries of the University of Auckland*

Description

A dataset containing the SDG queries of University of Auckland (version 1). The queries are available from https://www.sdgmapping.auckland.ac.nz/. The Auckland queries were developed to build on the processes developed by the United Nations and the Times Higher Education Ranking in order to create an expanded list of keywords that can be used to identify SDG-relevant research. There is one query per SDG. There are no queries for SDG-17.

Usage

```
auckland_queries
```

Format

A data frame with 16 rows and 4 columns

system Name of system

sdg Label of the SDG

query_id Index of the query

query SDG query

Source

https://www.sdgmapping.auckland.ac.nz/

aurora_queries *SDG queries of the Aurora Universities Network*

Description

A dataset containing the SDG queries version 5.0 of the Aurora Universities Network. See the corresponding GitHub repository. For the actual implementation of the queries see `aurora_simple`, `aurora_and`, `aurora_w`, and the queries hard-coded in `detect_aurora`. There are multiple queries per SDG (one per row). In comparison to previous versions, this version of the queries Aurora added more keywords related to academic terminology to be able to detect more research papers related to the SDGs. The current version also drew inspiration from the SIRIS query system (`siris_queries`). The Aurora queries were designed to be precise rather than sensitive. To achieve this the queries make use complex keyword-combinations using several different logical search operators. All SDGs (1-17) are covered.

Usage

```
aurora_queries
```

Format

A data frame with 378 rows and 5 columns

system Name of system

sdg Label of the SDG

sdg_title Title of the SDG

sdg_description Description of the SDG

query_id Index of the query

query Original SDG query

Source

https://github.com/Aurora-Network-Global/sdg-queries/releases/tag/v5.0

crosstab_sdg *Compare query systems and SDGs*

Description

`crosstab_sdg` calculates cross tables (aka contingency tables) of SGSs or systems across hits identified via detect_sdg_systems.

Usage

```
crosstab_sdg(hits, compare = c("systems", "sdgs"), systems = NULL, sdgs = NULL)
```

Arguments

hits	data frame as returned by detect_sdg_systems. Must include columns document, sdg, system, and hit.
compare	character specifying whether systems or SDGs should be cross tabulated.
systems	character vector specifying the query systems to be cross tabulated. Values must be available in the system column of hits. Defaults to NULL in which case available values are retrieved from hits.
sdgs	numeric vector with integers between 1 and 17 specifying the SDGs to be cross tabluated. Values must be available in the sdg column of hits. Defaults to NULL in which case available values are retrieved from hits.

Details

crosstab_sdg determines correlations between either query systems or SDGs. The respectively other dimension will be ignored. Note that correlations between SDGs may vary between query systems.

Value

matrix showing correlation coefficients for all pairs of query systems (if compare = "systems") or SDGs (if compare = "SDGs").

Examples

```
# run sdg detection
hits <- detect_sdg_systems(projects)

# create cross table of systems
crosstab_sdg(hits)

# create cross table of systems
crosstab_sdg(hits, compare = "sdgs")
```

---

detect_any *Detect SDGs in text with own query system*

Description

detect_any identifies SDGs in text using user provided query systems. Works like detect_sdg_systems but uses a user specified query system instead of an existing one like detect_sdg_systems does.

Usage

```
detect_any(
  text,
  system,
  queries = lifecycle::deprecated(),
  sdgs = NULL,
  output = c("features", "documents"),
  verbose = TRUE
)
```

Arguments

`text`	`character` vector or object of class `tCorpus` containing text in which SDGs shall be detected.
`system`	a data frame that must contain the following variables: a `character` vector with queries, a `integer` vector specifying which SDG each query maps to (values must be between 1 and 17) and a `character` with one unique value specifying the name of the used query system (can be anything as long as it is unique).
`queries`	deprecated.
`sdgs`	`numeric` vector with integers between 1 and 17 specifying the sdgs to identify in `text`. Defaults to `1:17`.
`output`	`character` specifying the level of detail in the output. The default `"features"` returns a `tibble` with one row per matched query, include a variable containing the features of the query that were matched in the text. By contrast, `"documents"` returns an aggregated `tibble` with one row per matched sdg, without information on the features.
`verbose`	`logical` specifying whether messages on the function's progress should be printed.

Value

The function returns a `tibble` containing the SDG hits found in the vector of documents. Depending on the value of `output` the tibble will contain all or some of the following columns:

document Index of the element in `text` where match was found. Formatted as a factor with the number of levels matching the original number of documents.

sdg Label of the SDG found in document.

systems The name of the query system that produced the match.

query_id Index of the query within the query system that produced the match.

features Concatenated list of words that caused the query to match.

hit Index of hit for a given system.

Examples

```
# create data frame with query system
my_queries <- tibble::tibble(
  system = "my_system",
  query = c(
    "theory",
    "analysis OR analyses OR analyzed",
    "study AND hypothesis"
  ),
  sdg = c(1, 2, 2)
)

# run sdg detection with own query system
hits <- detect_any(projects, my_queries)

# run sdg detection for sdg 2 only
hits <- detect_any(projects, my_queries, sdgs = 2)
```

`detect_sdg` *Detect SDGs in text using ensemble model*

Description

`detect_sdg` identifies SDGs in text using an ensemble model approach considering multiple existing SDG query systems and text length.

Usage

```
detect_sdg(
  text,
  systems = lifecycle::deprecated(),
  output = lifecycle::deprecated(),
  sdgs = 1:17,
  synthetic = c("equal"),
  verbose = TRUE
)
```

Arguments

`text`	`character` vector or object of class `tCorpus` containing text in which SDGs shall be detected.
`systems`	As of text2sdg 1.0.0 the 'systems' argument of 'detect_sdg()' is deprecated. This is because 'detect_sdg()' now makes use of an ensemble approach that draws on all systems as well as on the text length, see –preprint– for more information. The old version of 'detect_sdg()' is available through the 'detect_sdg_systems()' function.

`output` As of text2sdg 1.0.0 the 'output' argument of 'detect_sdg()' is deprecated. This is because 'detect_sdg()' now makes use of an ensemble approach that draws on all systems as well as on the text length, see –preprint– for more information. The old version of 'detect_sdg()' is available through the 'detect_sdg_systems()' function.

`sdgs` `numeric` vector with integers between 1 and 17 specifying the sdgs to identify in `text`. Defaults to `1:17`.

`synthetic` `character` vector specifying the ensemble version to be used. These versions vary in terms of the amount of synthetic data used in training (relative to the amount of expert-labeled data). Can be one or more of `"none"`, `"third"`, `"equal"`, and `"triple"`. The default is `"equal"`.

`verbose` `logical` specifying whether messages on the function's progress should be printed.

Details

`detect_sdg` implements a ensemble model to detect SDGs in text. The ensemble model combines the six systems implemented by `detect_sdg_systems` and text length in a random forest architecture. The ensemble model has been trained on three data sets with SDG labels assigned by experts and a matching number of synthetic texts generated by random sampling from a word frequency list. The user has the choice of multiple versions of the ensemble model that have been trained on different amounts of synthetic texts to adjust the sensitivity and specificity of the model. Increasing the amount of of synthetic data makes the ensemble more conservative, leading to increased sensitivity and decreased specificity.

By default, `detect_sdg` implements the version of the ensemble model that has been trained on an equal amount of expert-labeled and synthetic data, providing a reasonable balance between sensitivity and specificity. For details, see article by Wulff et al. (2023).

Value

The function returns a `tibble` containing the SDG hits found in the vector of documents. The columns of the `tibble` are described below. The `tibble` also includes as an attribute with name `"system_hits"` the predictions of the individual systems produced by `detect_sdg_systems()`.

document Index of the element in `text` where match was found. Formatted as a factor with the number of levels matching the original number of documents.

sdg Label of the SDG found in document.

system The name of the ensemble system that produced the match.

hit Index of hit for the Ensemble model.

References

Wulff, D. U., Meier, D., & Mata, R. (2023). Using novel data and ensemble models to improve automated SDG-labeling. arXiv

Examples

```
# run sdg detection
hits <- detect_sdg(projects)

# run sdg detection for sdg 3 only
hits <- detect_sdg(projects, sdgs = 3)

# extract systems hits
attr(hits, "system_hits")
```

`detect_sdg_systems` *Detect SDGs in text*

Description

`detect_sdg_systems` identifies SDGs in text using multiple SDG query systems.

Usage

```
detect_sdg_systems(
  text,
  systems = c("Aurora", "Elsevier", "Auckland", "SIRIS"),
  sdgs = 1:17,
  output = c("features", "documents"),
  verbose = TRUE
)
```

Arguments

`text`	`character` vector or object of class `tCorpus` containing text in which SDGs shall be detected.
`systems`	`character` vector specifying the query systems to be used. Can be one or more of `"Aurora"`, `"Elsevier"`, `"Auckland"`, `"SIRIS"` `"SDSN"`, and `"SDGO"`. By default all systems except `"SDGO"` and `"SDSN"` are used.
`sdgs`	`numeric` vector with integers between 1 and 17 specifying the sdgs to identify in `text`. Defaults to `1:17`.
`output`	`character` specifying the level of detail in the output. The default `"features"` returns a `tibble` with one row per matched query, include a variable containing the features of the query that were matched in the text. By contrast, `"documents"` returns an aggregated `tibble` with one row per matched sdg, without information on the features.
`verbose`	`logical` specifying whether messages on the function's progress should be printed.

Details

detect_sdg_systems implements six SDG query systems. Four systems developed by the Aurora Universities Network (see aurora_queries), Elsevier (see elsevier_queries), Auckland University (see elsevier_queries), and SIRIS Academic (see siris_queries) rely on Lucene-style Boolean queries, whereas two systems, namely SDGO (see sdgo_queries) and SDSN (see sdsn_queries) rely on basic keyword matching. 'detect_sdg_systems' calls dedicated detect_* for each of the five system. Search of the queries is implemented using the search_features function from the corpustools package.

By default, detect_sdg_systems runs only the Aurora, Elsevier, Auckland, and Siris query systems, as they are considerably less liberal than the SDSN and SDGO systems and therefore likely produce more valid SDG classifications. Users should be aware that systematic validations and comparison between the systems are largely lacking and that results should be interpreted with caution.

Value

The function returns a tibble containing the SDG hits found in the vector of documents. The columns of the tibble depend on the value of output. Possible columns are:

document Index of the element in text where match was found. Formatted as a factor with the number of levels matching the original number of documents.

sdg Label of the SDG found in document.

system The name of the query system that produced the match.

query_id Index of the query within the query system that produced the match.

features Concatenated list of words that caused the query to match.

hit Index of hit for a given system.

n_hits Number of queries that produced a hit for a given system, sdg, and document.

Examples

```
# run sdg detection
hits <- detect_sdg_systems(projects)

# run sdg detection with Aurora only
hits <- detect_sdg_systems(projects, systems = "Aurora")

# run sdg detection for sdg 3 only
hits <- detect_sdg_systems(projects, sdgs = 3)
```

elsevier_queries *SDG queries of Elsevier*

Description

A dataset containing the SDG queries of Elsevier (version 1). The queries are available from data.mendeley.com. The Elsevier queries were developed to maximize SDG hits on the Scopus database. A detailed description of how each SDG query was developed can be found here. There is one query per SDG. There are no queries for SDG-17.

Usage

```
elsevier_queries
```

Format

A data frame with 16 rows and 4 columns

system Name of system

sdg Label of the SDG

query_id Index of the query

query SDG query

Source

https://data.mendeley.com/datasets/87txkw7khs/1

plot_sdg *Plot distributions of SDGs identified in text*

Description

`plot_sdg` creates a (stacked) barplot of the frequency distribution of SDGs identified via detect_sdg or detect_sdg_systems.

Usage

```
plot_sdg(
  hits,
  systems = NULL,
  sdgs = NULL,
  normalize = "none",
  color = "unibas",
  sdg_titles = FALSE,
  remove_duplicates = TRUE,
  ...
)
```

Arguments

hits	data frame as returned by detect_sdg or detect_sdg_systems. Must include columns sdg and system.
systems	character vector specifying the query systems to be visualized. Values must be available in the system column of hits. systems of length greater 1 result, by default, in a stacked barplot. Defaults to NULL in which case available values are retrieved from hits.
sdgs	numeric vector with integers between 1 and 17 specifying the SDGs to be visualized. Values must be available in the sdg column of hits. Defaults to NULL in which case available values are retrieved from hits.
normalize	character specifying whether results should be presented as frequencies (normalize = "none"), the default, or whether the frequencies should be normalized using either the total frequencies of each system (normalize = "systems") or the total number of documents (normalize = "documents").
color	character vector used to color the bars according to systems. The default, "unibas", uses three colors of University of Basel's corporate design. Alternatively, color must specified using color names or color hex values. color will be interpolated to match the length of systems.
sdg_titles	logical specifying whether the titles of the SDG should added to the axis annotation.
remove_duplicates	logical specifying the handling of multiple hits of the same SDG for a given document and system. Defaults to TRUE implying that no more than one hit is counted per SDG, system, and document.
...	arguments passed to geom_bar.

Details

The function is built using ggplot and can thus be flexibly extended. See examples.

Value

The function returns a ggplot object that can either be stored in an object or printed to produce the plot.

Examples

```
# run sdg detection
hits <- detect_sdg_systems(projects)

# create barplot
plot_sdg(hits)

# create barplot with facets
plot_sdg(hits) + ggplot2::facet_wrap(~system)
```

projects *Descriptions of research projects*

Description

500 project descriptions of University of Basel research projects that were funded by the Swiss National Science Foundation. The project descriptions were drawn randomly from University of Basel projects listed in the the public P3 project data base.

Usage

```
projects
```

Format

A `character` vector of length 500.

Source

https://data.snf.ch/about/glossary

sdgo_queries *SDG Ontology by OSDG*

Description

A dataset containing the SDG queries based on the keyword ontology by OSDG. The queries are available from figshare.com.

Usage

```
sdgo_queries
```

Format

A data frame with 4,122 rows and 5 columns

system Name of system

sdg Label of the SDG

keyword SDG keyword used in query

query_id Index of the query

query SDG query

Details

Bautista-Puig, N.; Mauleón E. (2019). Unveiling the path towards sustainability: is there a research interest on sustainable goals? In the 17th International Conference on Scientometrics & Informetrics (ISSI 2019), Rome (Italy), Volume II, ISBN 978-88-3381-118-5, p.2770-2771. The authors of these queries first created an ontology from central keywords in the SDG UN description and expanded these keywords with keywords they identified in SDG related research output. There are multiple queries per SDG. All SDGs (1-17) are covered.

Source

https://figshare.com/articles/dataset/SDG_ontology/11106113/1

---

sdsn_queries *SDG keywords by SDSN*

Description

A dataset containing SDG-specific keywords compiled from several universities from the Sustainable Development Solutions Network (SDSN) Australia, New Zealand & Pacific Network. The authors used UN documents, Google searches and personal communications as sources for the keywords. All SDGs (1-17) are covered.

Usage

```
sdsn_queries
```

Format

A data frame with 847 rows and 5 columns

system Name of system

sdg Label of the SDG

keyword SDG keyword used in query

query_id Index of the query

query SDG query

Source

https://ap-unsdsn.org/regional-initiatives/universities-sdgs/

siris_queries *SDG queries of SIRIS Academic*

Description

A dataset containing the SDG queries of SIRIS Academic. The queries are available fromZenodo.org. The SIRIS queries were developed by extracting key terms from the UN official list of goals, targets and indicators as well from relevant literature around SDGs. The query system has subsequently been expanded with a pre-trained word2vec model and an algorithm that selects related words from Wikipedia. There are multiple queries per SDG (one per row). There are no queries for SDG-17.

Usage

```
siris_queries
```

Format

A data frame with 3,445 rows and 6 columns

system Name of system

sdg Label of the SDG

keyword Primary SDG query element

extra Secodary SDG query element

query_id Index of the query

query SDG query

Source

https://zenodo.org/record/3567769#.YVMhH9gzYUG

text2sdg *Detecting UN Sustainable Development Goals in Text*

Description

The text2sdg package provides functions for detecting SDGs in text, as well as for analyzing and visualization the hits found in text. The following provides a brief overview of the contents of the package.

Detect functions

`detect_sdg` detects SDGs in text using up to five different query systems: Aurora, Elsevier, SIRIS, SDSN, and OSDG

`detect_any` detects SDGs in text using self-specified queries utilizing the lucene-style syntax of the `corpustools` package.

Analysis functions

plot_sdg visualizes the relative frequency of SDG hits across query systems.

crosstab_sdg calculates cross tables of correlations between either the query systems or the different SDGs.

Datasets

projects contain random selection of research project descriptions from the P3 database of the Swiss National Science Foundation.

aurora_queries, elsevier_queries, siris_queries, sdsn_queries, auckland_queries and sdgo_queries contain a mapping of SDGs and search queries as they are employed in the respective systems.

Examples

```
# detect SDGs using default systems
hits <- detect_sdg_systems(projects)

#' # detect SDGs using all five systems
hits <- detect_sdg_systems(projects,
  system = c("Aurora", "Elsevier", "SIRIS", "SDSN", "SDGO")
)

# visualize SDG frequencies
plot_sdg(hits)

# correlations between systems
crosstab_sdg(hits)

# correlations between SDGs
crosstab_sdg(hits, compare = "sdgs")
```

Index