Package: openalexSnowball (via r-universe)

June 2, 2026

Type Package

Title Snowball searches for OpenAlex using the openalexPro pipeline

Version 0.1.4

Author Rainer M Krug

Maintainer Rainer M Krug <Rainer@krugs.de>

Description Perform snowball searches on the OpenAlex citation graph using openalexPro's on-disk processing pipeline and store results in Parquet.

URL https://github.com/openalexPro/openalexSnowball, https://openalexpro.github.io/openalexSnowball/, https://doi.org/10.5281/zenodo.20448982

BugReports https://github.com/openalexPro/openalexSnowball/issues

License GPL (>= 2)

Depends R (>= 4.1.1)

Imports openalexPro (>= 0.4.2), arrow, DBI, dplyr, duckdb, rlang

Additional_repositories https://openalexpro.r-universe.dev

Suggests keyring, knitr, openalexR (>= 1.4.0), rmarkdown, quarto, testthat (>= 3.0.0), vcr (> 1.7.0), vdiffr

Encoding UTF-8

RoxygenNote 7.3.3

VignetteBuilder quarto

Config/testthat/edition 3

Config/pak/sysreqs cmake libjq-dev libssl-dev xz-utils

Repository https://openalexpro.r-universe.dev

Date/Publication 2026-06-02 15:39:46 UTC

RemoteUrl https://github.com/openalexPro/openalexSnowball

RemoteRef main

RemoteSha 7ffa641165bee2c19c29c6aacb4ddf52da595ac8

Contents

pro_snowball 2
pro_snowball_extract_edges 3
pro_snowball_get_nodes 4
read_snowball 5

Index **6**

`pro_snowball` *A function to perform a snowball search and convert the result to a tibble/data frame.*

Description

A function to perform a snowball search and convert the result to a tibble/data frame.

Usage

```
pro_snowball(
  identifier = NULL,
  doi = NULL,
  output = tempfile(fileext = ".snowball"),
  verbose = FALSE
)
```

Arguments

`identifier`	Character vector of openalex identifiers.
`doi`	Character vector of dois.
`output`	parquet dataset; default: temporary directory.
`verbose`	Logical indicating whether to show a verbose information. Defaults to `FALSE`

Value

The folder of the results containing multiple subfolders.

`pro_snowball_extract_edges` *A function to extract the edges from a parquet database containing the nodes*

Description

A function to extract the edges from a parquet database containing the nodes

Usage

```
pro_snowball_extract_edges(
  nodes = NULL,
  output = tempfile(fileext = ".snowball"),
  verbose = FALSE
)
```

Arguments

`nodes`	Path to the nodes parquet dataset
`output`	output folder, in which the parquet database containing the edges called `edges` will be savedp default: temporary directory.
`verbose`	Logical indicating whether to show a verbose information. Defaults to `FALSE`

Value

A list containing 2 elements:

- nodes: dataframe with publication records. The last column `oa_input` indicates whether the work was one of the input `identifier`(s).
- edges: publication link dataframe of 2 columns `from, to` such that a row `A, B` means A -> B means A cites B. In bibliometrics, the "citation action" comes from A to B.

Examples

```
## Not run:

snowball_docs <- pro_snowball(
   identifier = c("W2741809807", "W2755950973"),
   citing_params = list(from_publication_date = "2022-01-01"),
   cited_by_params = list(),
   verbose = TRUE
)

# Identical to above, but searches using paper DOIs

snowball_docs_doi <- oa_snowball(
   doi = c("10.1016/j.joi.2017.08.007", "10.7717/peerj.4375"),
```

```
  citing_params = list(from_publication_date = "2022-01-01"),
  cited_by_params = list(),
  verbose = TRUE
)

## End(Not run)
```

`pro_snowball_get_nodes` *A function to get the nodes for a snowball search*

Description

A function to get the nodes for a snowball search

Usage

```
pro_snowball_get_nodes(
  identifier = NULL,
  doi = NULL,
  limit = NULL,
  output = tempfile(fileext = ".snowball"),
  verbose = FALSE
)
```

Arguments

`identifier`	Character vector of openalex identifiers.
`doi`	Character vector of dois.
`limit`	If `citedOnly` only works cited by the keypaper are retrieved, `citingOnly` retrieves only works citing the keypaper. Default: `NULL` where all will be retrieved. 'none' is equal to `NULL`
`output`	parquet dataset; default: temporary directory.
`verbose`	Logical indicating whether to show a verbose information. Defaults to `FALSE`

Value

Path to the nodes parquet dataset

read_snowball *Read snowball from Parquet Dataset*

Description

This function reads a snowball from Apache Parquet format and returns a list containing nodes and edges, which can be either Arrow Datasets or tibbles.

Usage

```
read_snowball(
  snowball = NULL,
  edge_type = c("core", "extended", "outside"),
  return_data = FALSE,
  shorten_ids = FALSE
)
```

Arguments

snowball	The directory of the Parquet files as poppulater by pro_snowball().
edge_type	type of the returned edges. Possible values are: • core: only edges from or to the keypapers are selected • extended, only edges between the nodes are selected (this includes core edges) • outside: only edges where either the from or the to is not in nodes multiple are allowed.
return_data	Logical indicating whether to return an ArrowObject representing the corpus (default) or a tibble containing the whole corpus shou,d be returned.
shorten_ids	If TRUE the ids will be shortened, i.e. the part https://openalex.org/ will be removed

Value

A list containing two elements: nodes and edges, which are either ArrowObject representing the corpus or tibbles containing the data.

Index

pro_snowball, 2
pro_snowball_extract_edges, 3
pro_snowball_get_nodes, 4

read_snowball, 5