SciBib

Release 0.1.0

Gaël Cousin

Jan 14, 2024

CONTENTS

CHAPTER

ONE

INDICES AND TABLES

- genindex
- modindex
- search

1.1 The SciBib Package

This package enables scientific bibliographical data recovery from an author's orcid id. The main goal is to collect bibtex entries for the authors works and to collect abstracts for theses works.

Bibtex collection works fine provided the author has an orcid record and the sought article is referenced there, see data_query.OrcidWork.bibtex.

Abstract retrieval can be performed using ArXiv's API if the article is on the Arxiv and the author associated her/his orcid id with her/his arxiv account, see data_query.AuthorData.work_summary_from_arxiv.

Another option is to get a (sometimes more up-to-date) abstract scraping the journal's website. In this case, some legal or technical obstructions might appear. However, a tool to try this technique is provided, namely the scrape_abstract method of our OrcidWork class.

Another useful feature is to use the doi of a work to build an url that leads to the article in the publisher's website. This can be obtained with OrcidWork.doi.

Other data sources could be added in the future, depending of the users' suggestions/pull requests.

1.2 The data_query module

This module defines two classes that allow to parse author data from Orcid and arxiv. These are AuthorData and OrcidWork.

`class scibib.data_query.OrcidWork`

Methods:

__init__(work_data)	Instantiate single work object.
scrape_abstract()	Scrape the work's summary from the editor/journal's site.

Attributes:

path	Orcid path to the data.
title	Work title.
doi	The Work's doi.
bibtex	Return the bibtex entry for self from source.

`__init__`(*work_data*)

Instantiate single work object.

Parameters

`work_data` (*`nested lists/dictionaries`*) – part of a loaded json data corresponding to a single work, as obtained from orcid's API.

`property path`

Orcid path to the data.

`property title`

Work title.

`property doi`

The Work's doi.

Returns

the doi.

Return type

str

`property bibtex`

Return the bibtex entry for self from source.

Parameters

- **`source`** (*`str, optional`*) – Equals 'doi'. Defaults to 'doi'.
- **`future.`** (*`Other sources might be available in the`*) –

`scrape_abstract`()

Scrape the work's summary from the editor/journal's site. Beware that you might need authorization from the editor/journal to use this functionality.

`class` `scibib.data_query.`**`AuthorData`**

A class to parse Orcid author entries.

Methods:

`__init__`(orcid_id)	Instantiator
`work_summary_from_arxiv`(orcid_work)	Match work with an arxiv entry to provide a summary.

Attributes:

orcid_record	The raw orcid record as a parsed json.
arxiv_record	The raw arxiv record as an atom feed.
articles	list of article entries in the author's Orcid entry.
orcid_id_is_on_arxiv	Check if the author associated his/her Arxiv with Orcid.
arxiv_summaries_dic	Return dict that maps arxiv_entries -> abstracts for the author.

__init__(*orcid_id*)

Instantiator

Parameters

orcid_id (*str*) – The author’s orcid id

property orcid_record

The raw orcid record as a parsed json.

Returns

The raw orcid record as a parsed json (using json.load).

Return type

list

property arxiv_record

The raw arxiv record as an atom feed.

property articles

list of article entries in the author’s Orcid entry.

Returns

list of article entries, formatted as OrcidWork instances.

Return type

list[*OrcidWork*]

property orcid_id_is_on_arxiv

Check if the author associated his/her Arxiv with Orcid.

Returns

True if yes, False if no!

Return type

bool

property arxiv_summaries_dic

Return dict that maps arxiv_entries -> abstracts for the author.

work_summary_from_arxiv(*orcid_work*)

Match work with an arxiv entry to provide a summary.

Parameters

orcid_work (OrcidWork) – the work that needs summary.

Returns

The guessed summary

Return type

str

1.3 The abstract_collector module

This module defines the main_paragraph function.

Functions:

main_paragraph(url)	From a web page, return the paragraph with the biggest length.

`scibib.abstract_collector.`**`main_paragraph`**(*url*)

From a web page, return the paragraph with the biggest length.

Parameters

url (*str*) – the url of the web page to treat.

PYTHON MODULE INDEX

INDEX