

A Vendor's View on Reproducibility: Datasets, Tools, & Partnerships

Jason E. Rollins, PhD
October, 2017

Datasets



DERWENT INNOVATION

Datasets

We provide custom Web of Science datasets for specific data challenges.

SIE: Characterize iSchool Research Territory via Scholarly Data

ICConference SIE

Characterize iSchool Research Territory via Scholarly Data

Description

In this SIE, participants are expected to characterize iSchool research territory in a larger scientific environment by leveraging novel scholarly data. For instance, by using information retrieval, data analysis, Bibliometrics, information visualization, data mining, etc. methods, participants can investigate and propose a number of interesting and novel questions.

Call for Papers

The proposed ideas are not necessarily restricted to the following exemplar topics:

- What are the most important research topics of iSchool?
- What are each individual iSchool's strengths and specializations?
- How iSchool researchers collaborate with other scholars? Do iSchool researchers prefer to collaborate with the scholars from other domains?
- In the past few years, which topics are getting increasingly popular in iSchool and which are decaying?
- Which iSchool topics are more likely contributed by external research communities?
- How can one effectively visualize iSchool territory in a larger context?

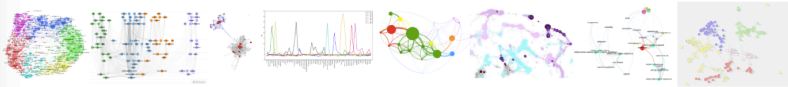
Important Data

- Online registration: TBD
- Dataset release: February 15, 2017
- Paper submission: March 20, 2017
- Announcement of results: TBD
- Conference presentation: March 24, 2017

Awards

TOPIC EXTRACTION CHALLENGE

[Home](#)
[Participate](#)
[The Data](#)
[Solutions](#)
[Publications](#)
[About](#)



We invite you to join the topic extraction challenge and learn about the state of art in topic extraction in bibliometrics through systematic comparison of topic extraction approaches applied by the various groups in the field and beyond. Over the last two years, six research teams worked together to compare their approaches to the identification of thematic structures in the Astronomy and Astrophysics literature, based on a shared set of bibliographic data of 111,616 journal articles. The outcomes of this comparative exercise are published in a forthcoming special issue of Scientometrics. Now that Clarivate Analytics kindly agreed to make this data set available to interested researchers in the bibliometrics community, we suggest to extend this comparative approach.

We invite you to participate in the comparative topic extraction challenge!

The challenge is not to develop the best partitioning of the data set. We believe this to be impossible because there is no single best solution for two reasons. First, the structure of a body of knowledge is in the eye of the beholder, i.e. more than one valid thematic structure can be constructed depending on the perspective applied to the knowledge. Second, topical structures are reconstructed for specific purposes, so if at all, there might be a best method for a given purpose. Therefore, we challenge you to use this opportunity to gain as much information as possible about your own approach and the reasons why it produced a particular solution, and to find out how it differs from solutions produced by other approaches. We challenge you to comparatively discuss advantages and disadvantages of approaches to topic identification and thus to contribute to a cumulative body of knowledge on the suitability of data models and algorithms for the identification of topics.

How to obtain the data set is described [here](#). Submitted solutions will be published [here](#) on this website (topic-challenge.info) and can be downloaded for comparisons. We will seek to make further tools available for comparison in the near future. If there are enough participants, we plan to run sessions on the comparative exercise at the next ISI conference and dedicated workshops. We hope that many of you will take up the challenge and thus contribute to cumulative research in bibliometrics.

<http://discern.uits.iu.edu:8826/sie/>

<http://www.topic-challenge.info>

Datasets

We make datasets available from our own research and work with customers under a variety of licensing options.

JOURNAL OF THE ASSOCIATION FOR
INFORMATION SCIENCE AND TECHNOLOGY

Open Access
 Creative Commons

RESEARCH ARTICLE

A Multidimensional Investigation of the Effects of Publication Retraction on Scholarly Impact

Xin Shuai, Jason Rollins, Isabelle Moulinier, Tonya Custis, Mathilda Edmunds, Frank Schilder

First published: 26 April 2017 [Full publication history](#)

DOI: 10.1002/asi.23826 [View/save citation](#)

Cited by (CrossRef): 0 articles [Check for updates](#) [Citation tools](#)

7

Abstract

During the past few decades, the rate of publication retractions has increased dramatically in academia. In this study, we investigate retractions from a quantitative perspective, aiming to answer two fundamental questions. One, how do retractions influence the scholarly impact of retracted papers, authors, and institutions? Two, does this influence propagate to the wider academic community through scholarly associations? Specifically, we analyzed a set of retracted articles indexed in Thomson Reuters Web of Science (WoS), and ran multiple experiments to compare changes in scholarly impact against a control set of nonretracted articles, authors, and institutions. We further applied the Granger Causality test to investigate whether different scientific topics are dynamically affected by retracted papers occurring within those topics. Our results show two key findings: first, the scholarly impact of retracted papers and authors significantly decreases after retraction, and the most severe impact decrease correlates with retractions based on proven, purposeful scientific misconduct; second, this retraction penalty does not seem to spread through the broader scholarly social graph, but instead has a limited and localized effect. Our findings may provide useful insights for scholars or science committees to evaluate the scholarly value of papers, authors, or institutions related to retractions.

[View issue TOC](#)
 Volume 68, Issue 9
 September 2017
 Pages 2225–2236

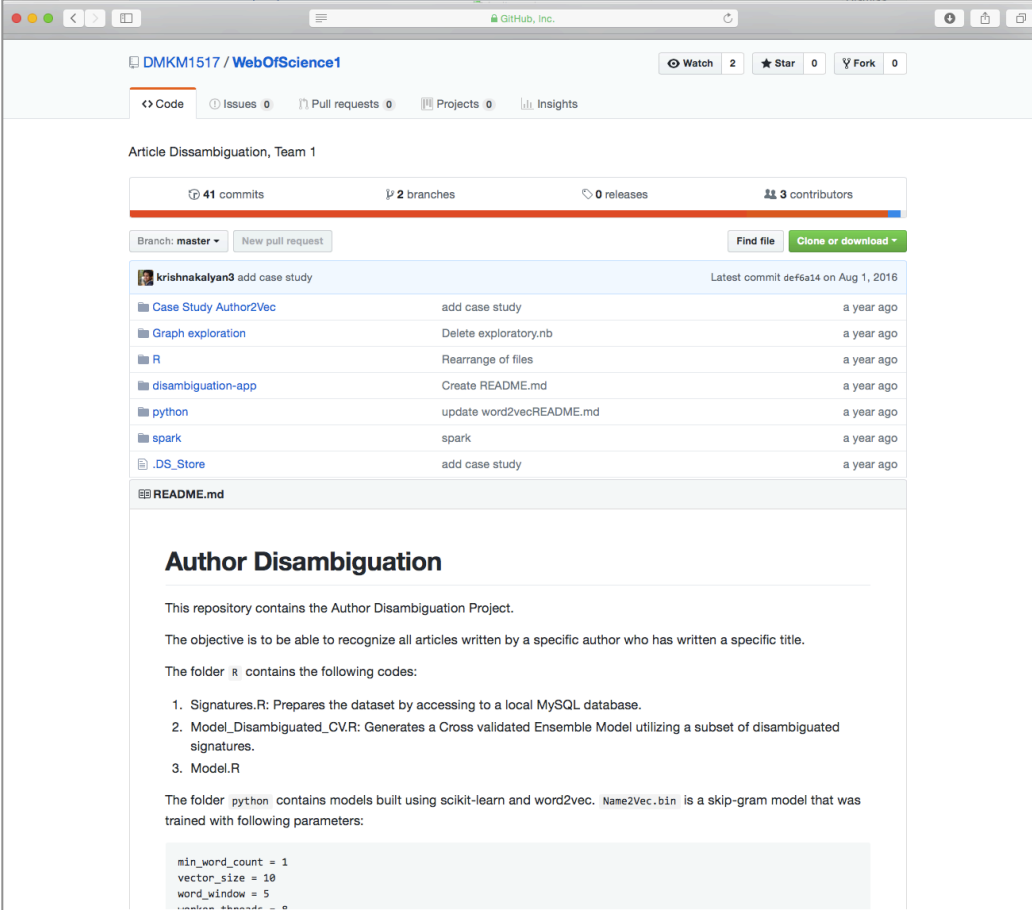
Tools

We publish recommendations on best practices.



Tools

We encourage researchers to share code and software optimized for WoS data.



The screenshot shows a GitHub repository page for 'DMKM1517 / WebOfScience1'. The repository has 41 commits, 2 branches, 0 releases, and 3 contributors. The latest commit is 'krishnakalyan3 add case study' from August 1, 2016. The repository contains several files and folders, including 'Case Study Author2Vec', 'Graph exploration', 'R', 'disambiguation-app', 'python', 'spark', and '.DS_Store'. The README.md file is visible, titled 'Author Disambiguation'. It describes the project's objective: to recognize all articles written by a specific author who has written a specific title. It lists the contents of the 'R' folder, including 'Signatures.R', 'Model_Disambiguated_CVR', and 'Model.R'. It also describes the 'python' folder, which contains models built using scikit-learn and word2vec. A code block at the bottom shows parameters for the word2vec model.

DMKM1517 / WebOfScience1

Watch 2 Star 0 Fork 0

Code Issues Pull requests Projects Insights

Article Disambiguation, Team 1

41 commits 2 branches 0 releases 3 contributors

Branch: master New pull request Find file Clone or download

krishnakalyan3 add case study Latest commit def6a14 on Aug 1, 2016

File/Folder	Description	Time
Case Study Author2Vec	add case study	a year ago
Graph exploration	Delete exploratory.nb	a year ago
R	Rearrange of files	a year ago
disambiguation-app	Create README.md	a year ago
python	update word2vecREADME.md	a year ago
spark	spark	a year ago
.DS_Store	add case study	a year ago

README.md

Author Disambiguation

This repository contains the Author Disambiguation Project.

The objective is to be able to recognize all articles written by a specific author who has written a specific title.

The folder `R` contains the following codes:

1. Signatures.R: Prepares the dataset by accessing to a local MySQL database.
2. Model_Disambiguated_CVR: Generates a Cross validated Ensemble Model utilizing a subset of disambiguated signatures.
3. Model.R

The folder `python` contains models built using scikit-learn and word2vec. `Name2Vec.bin` is a skip-gram model that was trained with following parameters:

```
min_word_count = 1
vector_size = 10
word_window = 5
num_epochs = 10
```

Partnerships – An Example...



- Meta-knowledge network: Researchers from Stanford, UCLA, Columbia, etc. Use WoS data for many publications – economic value of science and innovation, and industry – university collaborations.
- Have WoS data in their Cloud Kotta enclave and help manage secure access to data subsets
- Host <https://lists.uchicago.edu/web/info/wos4research>

