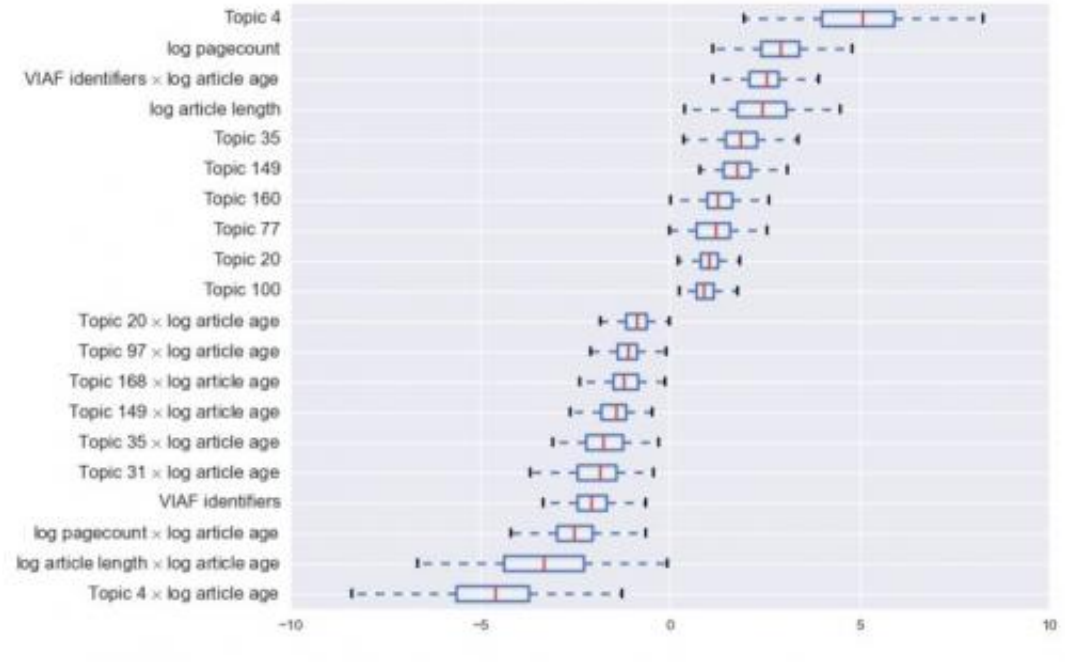
Algorithm, not live committee, performs author ranking

November 21 2014, by Nancy Owano



Coefficients of the regression of the presence of a digital edition on selected features. Credit: arXiv:1411.2180 [cs.DL]

Thousands of authors' works enter the public domain each year, but only a small number of them end up being widely available. So how to choose the ones taking center-stage? And how well can a machine-learning algorithm rank the most notable authors in the world? Allen B. Riddell at Dartmouth College set out to deliver some answers and he published his work, "Public Domain Rank: Identifying Notable Individuals with the

Wisdom of the Crowd", on the *ArXiv* server.

Riddell recognizes that identifying literary, scientific, and technical works of enduring interest is challenging. MIT Technology Review, reporting on emerging technology from the *ArXiv*, similarly noted how deciding which books to digitize when they enter the public domain is tricky unless you have an independent ranking of the most notable authors. Riddell's paper introduces an automatic method for identifying authors of notable works throughout history. In his paper, Riddell observed that "We have the empirical record of what works volunteers have edited and published in online repositories such as Project Gutenberg. In the deliberations of these volunteers, we have a valuable independent judgment of which works (and, by extension, which authors) have 'stood the test of time.' Unfortunately, this judgment is only reliable for works that have been in the public domain for a considerable amount of time; the collective judgment of the crowd is unavailable for works still covered by copyright monopolies."

The MIT Technology Review explanation of his approach: He uses a machine-learning algorithm to mine two databases. The first involves over 1 million online books in the public domain maintained by the University of Pennsylvania. The second is Wikipedia."

He makes use of Wikipedia entries of all authors in the English language edition. The algorithm extracts information such as article length, age and estimated views per day. The algorithm takes the list of all authors on the online book database and looks for a correlation between the biographical details on Wikipedia and the existence of a digital edition in the public domain. That produces a "public domain ranking" of all the authors that appear on Wikipedia.

Said the author: "This bottom-up approach to identifying in which works and individuals there is enduring interest makes use of two sources of

open data, a database of digital editions on the Online Books Page and Wikipedia. By aligning bibliographic records in the Online Books Page with the streams of structured and unstructured data from Wikipedia, this project facilitates the identification of notable works in or entering the public domain."

His paper presents an interesting point for debate, which has philosophical if not other touchpoints. In determining great works, does one feel more comfortable with decisions by select committees of expert elders? Does one actually prefer some subjectivity in the mix?

MIT Technology Review said, "The beauty of this approach is that it is entirely independent. That's in stark contrast to the committees that are often set up to rank works subjectively." Jon Fingas of Engadget thought "it's arguably an easier way to pick literary greats than leaving things up to an academic committee."

According to MIT Technology Review, Riddell said his ranking system compared well with existing rankings compiled by human experts, such as one compiled by the editorial board of the Modern Library.

More information: Public Domain Rank: Identifying Notable Individuals with the Wisdom of the Crowd, arXiv:1411.2180 [cs.DL] arxiv.org/abs/1411.2180

Abstract
Identifying literary, scientific, and technical works of enduring interest is challenging. Few are able to name significant works across more than a handful of domains or languages. This paper introduces an automatic method for identifying authors of notable works throughout history. Notability is defined using the record of which works volunteers have made available in public domain digital editions. A significant benefit of this bottom-up approach is that it also provides a novel and reproducible

index of notability for all individuals with Wikipedia pages. The method promises to supplement the work of cultural organizations and institutions seeking to publicize the availability of notable works and prioritize works for preservation and digitization.