KULA KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Meneses, Luis, et al. 2019. Aligning Social Media Indicators with the Documents in an Open Access Repository. *KULA: knowledge creation, dissemination, and preservation studies* 3(1): 19. DOI: https://doi.org/10.5334/kula.44

RESEARCH ARTICLE

Aligning Social Media Indicators with the Documents in an Open Access Repository

Luis Meneses[1], Alyssa Arbuckle[1], Hector Lopez[1], Belaid Moa[2], Richard Furuta[3] and Ray Siemens[1]

[1] Electronic Textual Cultures Lab, University of Victoria, CA

[2] University of Victoria, CA

[3] Center for the Study of Digital Libraries, Texas A&M University, US

Corresponding author: Luis Meneses (ldmm@uvic.ca)

In this paper we describe our current efforts towards building a framework that extends the functionality of an Open Access Repository by implementing processes to incorporate the ongoing trends in social media into the context of a digital collection. We refer to these processes collectively as the Social Media Engine. The purpose of our framework is twofold: first, we propose to challenge some of the preconceived notions of digital libraries by making repositories more dynamic; and second, by challenging this notion we want to promote public engagement and open scholarship. As a work in progress, we believe that a real challenge lies in investigating the implications that these two points introduce within the context of the humanities.

Keywords: social media; open access repositories; digital collections; user engagement

Introduction

In his pioneering 1945 essay *As We May Think*, Vannevar Bush envisions a time in which the world's knowledge is accessible by machine and in which the connections that describe the higher level relationships among sources are themselves objects of scholarship that can be shared with colleagues. We can see this today on the Web with the utility of resource lists such as Yahoo, the investigation of mechanisms such as Walden's Paths (Bogen et al. 2011) and the development of publishing platforms like the Public Knowledge Project (Owen and Stranack 2012). This is a natural side effect of collaboration and cooperation; as the problems to be solved in the humanities grow beyond the technical abilities of an individual scholar, and as social media becomes more embedded into our work practices, the presence of resources that situate knowledge within the broader environment will also become more prevalent.

Currently, the methods for representing documents and disseminating knowledge are changing. We have witnessed an increase in social media on the Web, which emphasizes its potential to transform the scholarly communication system (Sugimoto et al. 2016) and allows the types of human connections that Bush envisioned actionable as data. More so, there are no mechanisms in place to incorporate social media into the workflow of a digital collection.

We propose to challenge this notion by introducing a Social Media Engine into the framework of an Open Access Repository. This engine and its underlying framework aim to promote public engagement, open social scholarship, and social knowledge creation by matching readers with publications. The framework relies on the gathering and analysis of corpora harvested, indexed, and rendered from open access and academic materials.

The fundamental concepts behind our framework and its Social Media Engine can be explained through a three-point use case scenario, which is illustrated in **Figure 1**. First, our framework yields a list of topics related to individual entries and articles in the corpus by applying textual analysis techniques and topic modeling (Blei, Ng & Jordan 2003). Second, our engine identifies trending papers on social media by looking

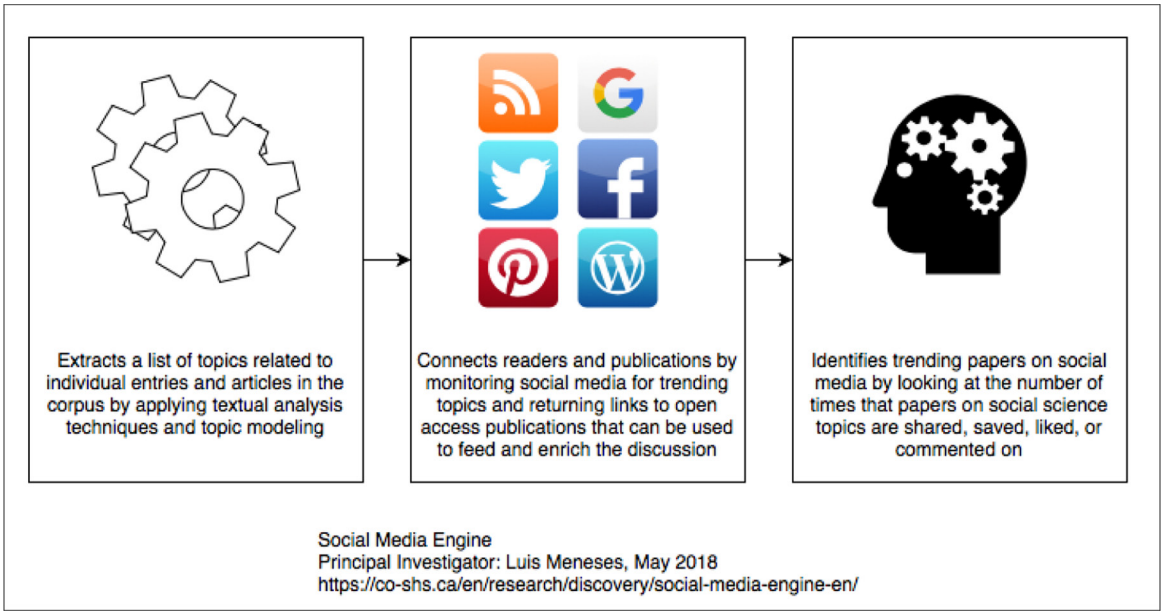

Figure 1: Fundamental concepts behind the Social Media Engine.

at the number of times that papers on social science topics are shared, saved, liked, or commented on. Finally, our engine connects readers and publications by monitoring social media for trending topics and returning links to open access publications that can be used to feed and enrich the discussion.

In this paper, we will elaborate on the findings obtained from implementing a prototype of our framework. The purpose of our study is twofold: first, we propose to challenge some of the preconceived notions of digital libraries by making repositories more dynamic; and second, by challenging this notion we want to promote public engagement and open scholarship. However, we believe that the real challenge lies in investigating implications of this work within the context of the humanities.

Background and Initial Analysis

We started our analysis by gathering an understanding of our document collection. Érudit.org is a digital repository of social sciences and humanities publications (Érudit Consortium 2017). This collection consists mostly of scholarly and cultural journals (100 and 30 respectively), theses, books, proceedings, and technical reports. The documents are diverse in content and span across 35 disciplines including arts, engineering, education, cinema, demography, law, theology, history, sociology, and women's studies, among others. As of February of 2017, the collection consists of 174,269 documents divided into 169 sets.

We proceeded to download the complete set of descriptive metadata through Érudit's Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) API. The OAI-PMH interface presents the documents as PDFs, which can be quite problematic to parse. Alternatively, a full text working set of the source materials were provided to us from Érudit's development platform. Since these documents were in XML format, we created a parser using Python's Beautiful Soup library (Richardson 2015) to pull the text from the markup elements. Using this parser, we found that 57 documents did not have full text associated with them—making their impact negligible when compared to the collection as a whole.

The majority of the articles in the repository are in French, and efforts to procure more articles in English by the Érudit Consortium are currently underway. To get a better sense of the document distribution, we ran a language detection process based on Google's language detection algorithms (Danilak 2017) on the full text articles. We found that 91% of the documents were in French, 8.6% in English and the remaining 0.4% in other languages. Why is this understanding of the corpus important? First, it helps us to examine our assumptions about the corpus; second, it allows us to grasp an overall understanding of the collection and set a foundation towards implementing solutions that can deal with documents in multiple languages; and third, it emphasizes the connections between different models and contributes towards the ongoing considerations of diversity in the digital humanities.

We then proceeded to further our initial understanding of the corpus from Érudit.org. More specifically, we were interested in how the descriptive metadata correlates and aligns with the full text contained in the documents. Upon analyzing the corpus, we found that 19% of the documents had metadata descriptions

and full text that were in the same language (i.e., metadata description and document text in French). Then we attempted to quantify the correlation between each description and its corresponding full text using three techniques: topic modeling with Latent Dirichlet allocation (LDA), term frequency–inverse document frequency (Tf-Idf), and resemblance with Jaccard coefficient. Out of the three, topic modeling and Tf-Idf had the best performance clustering the full text with the description in most of the cases.

System Architecture

Our framework consists of three main components which are hosted under Compute Canada's Westgrid server infrastructure (Westgrid 2017): 1) a keyword extraction module, 2) a social media mining component and 3) a search engine. These components and their corresponding interactions are illustrated in **Figure 2**, and will be described in more detail in the following sections of this paper.

Keyword Extraction: Search Engine Alignment

After our initial analysis and definition of the system architecture, we started to include topic modeling into our analysis. We performed topic modeling using the implementation of LDA available in Gensim (Řehůřek 2017) by dividing the corpus into its different OAI-PMH sets and checking if the results of the analysis align with our overall understanding of the corpus. As a part of this stage of the study, we worked on retrieving documents from the corpus using the terms that were unique for each topic. For testing purposes, we proceeded to increase the granularity of the sample and used documents from an OAI-PMH set labeled 'rabaska96'; this set contained 792 documents exclusively in French. We modeled these documents into 15 topics and extracted the terms that were unique for each topic.

We used Whoosh (Chaput 2017), a search engine that is built with Python (van Rossum 1995), to test if the documents from each topic could be retrieved using their identifying features, which in this case corresponds to their unique terms. We found that the documents in the set were being matched using only the unique

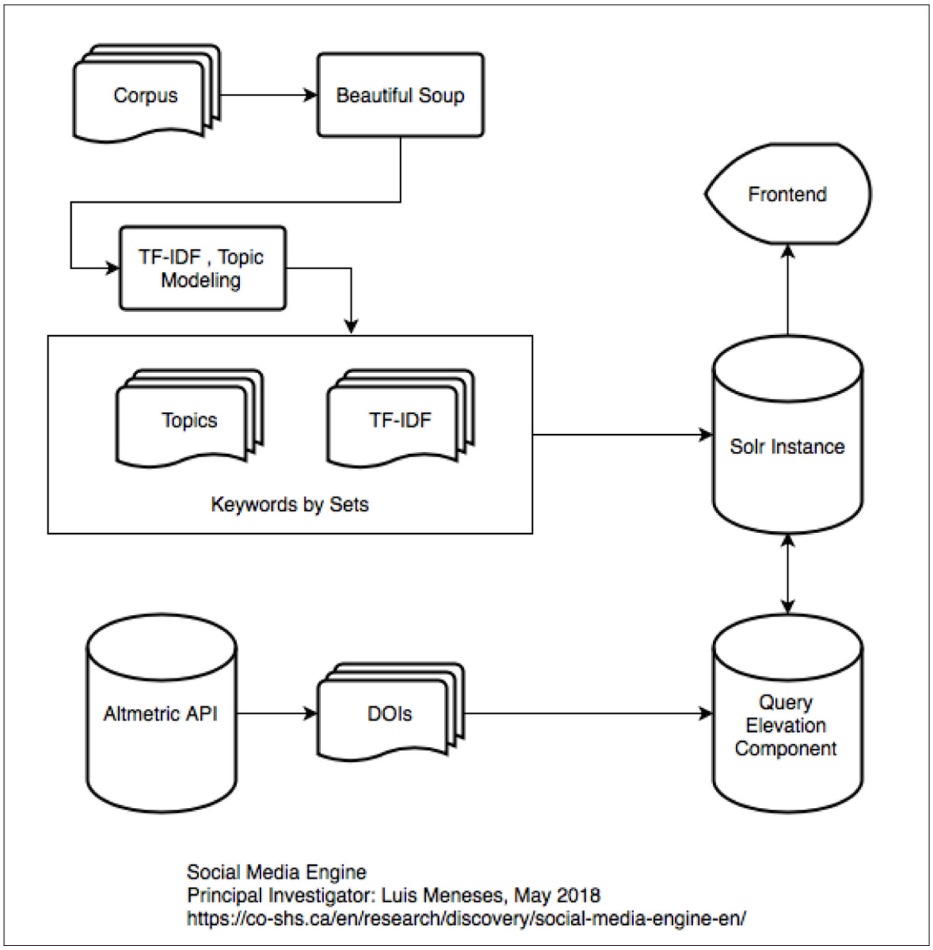

Figure 2: System diagram for the Social Media Engine.

terms gathered from the topic modeling. Furthermore, these findings support our hypothesis wherein unique terms extracted from each topic can be used to boost certain facets of the query. Consequently, we hypothesized at this point that these unique terms will facilitate the re-ranking of the results.

Modeling the Corpus

During this phase, we also started experimenting with parallel computing solutions based on Apache Spark (Apache Software Foundation 2017a) and its implementation of LDA clustering. The use of parallelization brought two advantages: first, the increase in speed is substantial; and second, the codebase is more robust—a byproduct of the optimization to afford parallelism.

The documents were ingested into a Spark Dataframe and preprocessed with five steps before undergoing the modeling stage. First, the text in each document was tokenized and lowercased. Second, tokens that contained non-alpha numeric characters were removed. Third, tokens that had less than four characters were removed as well. Fourth, we removed proper names from the corpus. To achieve this, we downloaded comma-separated values (CSV) files containing first names (both male and female) and last names from the United States of America census site from 1990 (US Census Bureau 2017). The files were then parsed and transformed into a reusable Python list with 91,910 entries. Finally, we created a list of stop words that consisted of 1/3 of the most common terms, which was appended to the proper names from step four. This preprocessing, along with multiple iterations per model, allowed us to obtain appropriate terms that reference the topics as results.

Taking advantage of the speed increase though parallelization, we were able to model the entire corpus, which we separated into sets to ensure the scalability of our approach. We modeled the 174,212 documents into 169 sets, and further clustered the documents into 20 topics per set with 100 iterations per model. For the most part, we were able to identify terms that can be linked specifically to each individual OAI-PMH set. The results from modeling the documents gave us a set of terms that can be used to cluster each set and a probability distribution for each document, which we saved as XML files for convenience. As described in the evaluation section of this paper, we used these terms and probability distributions to perform a simple evaluation.

Mining Social Media

Connecting readers with publications and monitoring social media to identify trending topics are fundamental aspects in our framework. However, these two points are complex research problems on their own. To streamline this complicated process and provide our framework with a greater level of abstraction, we decided to use Altmetric.com (Altmetric LLP 2017) for our social media mining. Almetric.com is a Web service that monitors several social media streams (including wikis, scholarly blogs, Twitter and Facebook), thereby helping scholars get a better overview of their scholarly content. More importantly, Almetric's API provides some level of granularity in their searching, including by date and DOI range. Furthermore, we used Pyaltmetric (Earp 2017) to query Almetric's API by creating and using sets of XML files that can be parsed into the search engine component of our prototype.

It is worthwhile to note the implications of using an external web service. The level of abstraction that we achieved can be seen as an advantage. On the other hand, we have no control over how Altmetric harvests and processes social media content and produces its results. We will expand on these points in the discussion section of this paper.

The Search Engine Component

We used Apache Solr (Apache Software Foundation 2017b) as the search engine for our framework. Solr is a robust, open source enterprise search platform that is based on Apache Lucene (Apache Software Foundation 2011). Ingesting the document corpus into Solr is done through XML files and by defining a schema, which defines the fields and metadata that will be used for searching. Given the robustness of Solr, ingesting Érudit's corpus into a core was not a difficult process: we had to specify an adequate configuration and find a directory location that could hold the document index. More importantly, Solr's Query Elevation Component allowed us to boost some search results depending on their relevance to mentions in social media: ranking is boosted using sets of keywords that match both social media and the documents in our corpus.

The biggest challenge in using Solr is finding an adequate front end. It is a very common practice to develop a custom front end for searching and browsing a collection, as several programming libraries for interfacing with Solr's API do exist. We decided to use a customized version of the Velocity Response Writer, a Solr front end based on Apache's Velocity Project (Apache Software Foundation 2017c), to provide us with the user interfacing functionality that we needed. **Figure 3** shows a screenshot of our user interface.

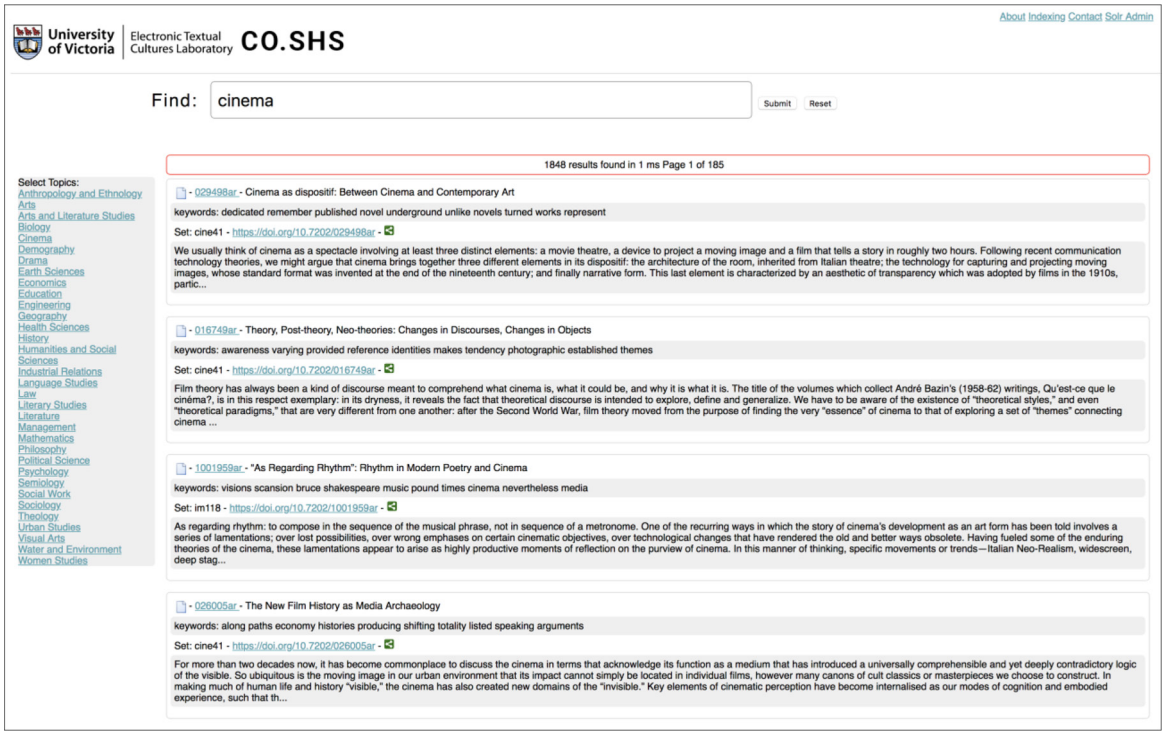

Figure 3: Screenshot of the Social Media Engine user interface.

Evaluation

We used a simple ranking function to evaluate our results. This ranking function is based on adding the Tf-Idf values to the documents, which were calculated using the terms obtained from the topic modeling for each set as a vocabulary. The output from this ranking was calculated for each set and, as before, the results for the ranking per set were saved in XML files.

In the end, we performed ranking calculations for three sets of terms: the first using the terms from the topic modeling; the second using terms gathered using Tf-Idf; and the third using a combination of topic modeling and Tf-Idf terms. In the third case, duplicated terms were removed to avoid overlap.

Using this simple ranking function, the Tf-Idf terms provided better rankings over the topic modeling terms. This was expected to some extent, given that the terms come from the documents themselves. On the other hand, the terms from the topic modeling also come from the documents, but they convey different information regarding the overall 'aboutness' of the collection. This 'aboutness' is important, especially since we are not aiming for higher values of precision (very specific to a formulated query), but rather higher recall (better document coverage for a given set of keywords). **Figure 4** provides a representation of the average rankings for each OAI-PMH set.

Ultimately, contextualizing the importance of the different modeling methods we used will let us outline the value and significance of our future evaluation steps—which include running a user study. However, a user study introduces a new layer of complexity, as elaborated on below.

Discussion

One of the main difficulties of the type of research that we outline in this paper is finding a way to evaluate the results obtained. We debated between the two obvious alternatives for evaluation: automated numerical methods versus running a user study with human subjects. Automated numerical methods are convenient, accurate, and inherently replicable. Their implied convenience comes with a cost, however: their accuracy can hinder the modeling of behaviors and thinking of human subjects. On the other hand, user studies can portray a better image of behaviors and user intentions—but they are both time and resource intensive. In our case, we decided to compromise and find a balance between the two alternatives. In the end, the automated numerical methods did provide us with the necessary feedback to lead us into the next stage of our research.

Although the terms obtained with Tf-Idf provided better results in the clustering by ranking, we do not discard the validity of the terms obtained with topic modeling. As stated earlier, we believe that these terms

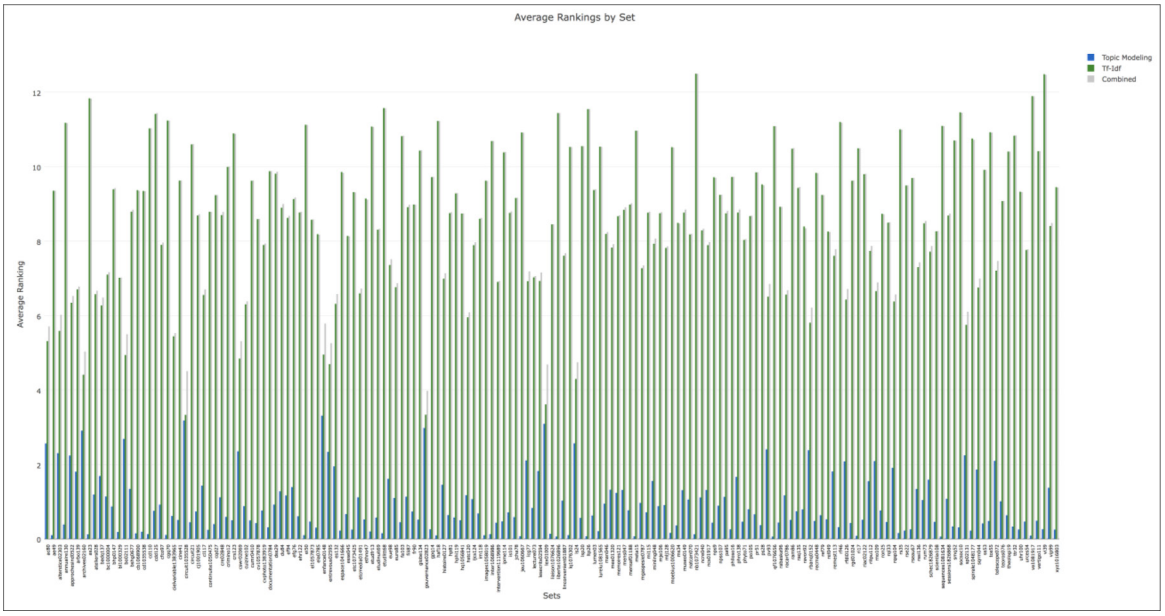

Figure 4: Average rankings by set.

provide an additional layer of metadata, which introduces the potential for an alternative method for their classification. More importantly, we believe that the combination of the terms obtained using Tf-Idf and topic modeling provides a layer of metadata that is very rich in its entropy—measured in regards to the amount of information that it conveys. More so, we hypothesize that this metadata layer can be used to dynamically reorganize a digital collection and align it with current trends in social media. We plan to explore this last hypothesis in future iterations of our work that involve modeling trends in social media and testing the alignment of these trends with our document corpus.

Another point that demands our attention is the use of external APIs for identifying trends in social media, as they overlook ongoing formulations within the digital humanities as sets of institutional infrastructures that exist within larger infrastructural ecologies (Liu 2016). Using Altmetric.com's API does streamline our approach, but identifying ongoing trends in social media is a complex problem in itself. Moreover, because of this complexity, it does fall outside the scope of our research. On the other hand, the level of abstraction that we sought forced us to use Altmetric as a 'black box': a method that produces output without our knowing its inner workings, potentially hindering our overall outcome if the API methods are flawed. However, Altmetric is considered a reputable source and offers other research institutions access to its API, which provides some assurance on the soundness of the metadata they provide. In an ideal scenario, we would have used a harvester developed by us or our research partners. Given that a custom solution does not exist, we decided to compromise.

Conclusions and future work

In this paper, we have outlined the findings from implementing a prototype for our system, which was concerned with aligning keywords with sets of documents using textual analysis techniques. We found that some techniques were more effective than others (notably Tf-Idf over topic modeling); however, their effectiveness is relative in this case. We concluded that one of the more complicated aspects of our study is its assessment. For this purpose, we have carried out automated testing, and two rounds of user studies are scheduled for the near future. Nonetheless, the results that we obtained do prove that sets of documents align with different sets of keywords. This has provided us with the necessary groundwork to match documents with ongoing trends in social media.

Through our research, we have developed a clear understanding of document collection and points of inflection to align a document's more prominent features with sets of keywords. Future work will focus on the evaluation of techniques and technologies for extracting features from social media. Further, we will perform test cases on the alignment of these social media indicators with the models that we have extracted from our document corpus. Ultimately, our research aims to build on Bush's vision and create tools to emphasize the connections between documents that can be treated as objects of study as well.

Competing Interests

The authors have no competing interests to declare.

References

Altmetric LLP. 2017. "Altmetric." Accessed October 20, 2017. https://www.altmetric.com/.

Apache Software Foundation. 2011. "Welcome to Apache Lucene." Accessed March 29, 2011. http://lucene.apache.org.

Apache Software Foundation. 2017a. "Apache Spark: Lightning-Fast Cluster Computing." Accessed April 11, 2017. http://spark.apache.org.

Apache Software Foundation. 2017b. "Apache Solr." Accessed October 20, 2017. http://lucene.apache.org/solr/.

Apache Software Foundation. 2017c. "The Apache Velocity Project." Accessed October 20, 2017. http://velocity.apache.org/.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research*, 3: 993–1022. http://www.jmlr.org/papers/v3/.

Bogen, Paul L., Daniel Pogue, Faryaneh Poursardar, Yuangling Li, Richard Furuta, and Frank Shipman. 2011. "WPv4: A Re-Imagined Walden's Paths to Support Diverse User Communities." In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 419–20. DOI: https://doi.org/10.1145/1998076.1998164

Bush, Vannevar. 1945. "As We May Think." *The Atlantic Monthly*, July 1945.

Chaput, Matt. 2017. "Whoosh: Fast Pure-Python Full Text Indexing, Search and Spell Checking Library." Accessed April 11, 2017. https://pypi.python.org/pypi/Whoosh/.

Danilak, Michal. 2017. "Langdetect 1.0.7." Accessed April 11, 2017. https://pypi.python.org/pypi/langdetect.

Earp, Will. 2017. *Pyaltmetric: Python Altmetric API v1 Wrapper* (version 0.2.0). OS Independent. Python.

Érudit Consortium. 2017. "Erudit.Org." Accessed April 11, 2017.

Liu, Alan. 2016. "Drafts for *Against the Cultural Singularity* (Book in Progress)." *Alan Liu*, May 2. http://liu.english.ucsb.edu/drafts-for-against-the-cultural-singularity/.

Owen, Brian, and Kevin Stranack. 2012. "The Public Knowledge Project and Open Journal Systems: Open Source Options for Small Publishers." *Learned Publishing*, 25(2): 138–44. DOI: https://doi.org/10.1087/20120208

Řehůřek, Radim. 2017. "Gensim: Topic Modelling for Humans." Accessed April 12, 2017. https://radimrehurek.com/gensim/.

Richardson, Leonard. 2015. "Beautiful Soup: We Called Him Tortoise Because He Taught Us." Accessed May 3, 2015. https://www.crummy.com/software/BeautifulSoup/.

Rossum, Guido van. 1995. "Python Tutorial, Technical Report CS-R9526." Amsterdam: Centrum voor Wikunde en Informatica (CWI). https://ir.cwi.nl/pub/5007/05007D.pdf.

Sugimoto, Cassidy R., Sam Work, Vincent Larivière, and Stefanie Haustein. 2016. "Scholarly Use of Social Media and Altmetrics: A Review of the Literature." *ArXiv:1608.08112 [Cs]*. http://arxiv.org/abs/1608.08112.

US Census Bureau. 2017. "Frequently Occurring Surnames from Census 1990 – Names Files." Accessed April 12, 2017. https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html.

Westgrid. 2017. "WestGrid: Compute Canada Regional Partner." Accessed April 11, 2017. https://www.westgrid.ca.

How to cite this article: Meneses, Luis, Alyssa Arbuckle, Hector Lopez, Belaid Moa, Richard Furuta and Ray Siemens. 2019. Aligning Social Media Indicators with the Documents in an Open Access Repository. *KULA: knowledge creation, dissemination, and preservation studies* 3(1): 19. DOI: https://doi.org/10.5334/kula.44

Submitted: 29 May 2018 **Accepted:** 07 September 2018 **Published:** 27 February 2019

]u[*KULA: knowledge creation, dissemination, and preservation studies* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS