POP
Public. Open. Participatory.

Pop! ▪ Issue N°1 ▪ **Dispersed/Networked Open Social Discovery Research**

Dispersed/Networked Open Social Discovery Research: Applications for Humanistic Machine Learning & Topic Modelling

RICHARD J. LANE

MeTA Digital Humanities Lab, Vancouver Island University – https://research.viu.ca/meta-digital-humanities-lab

OCTOBER 31, 2019

Introduction

One of the benefits of open social scholarship also presents researchers with a challenge: the *dispersed nature* of the knowledge breakthroughs presented by a diverse network of scholars inside and outside of the academy, or what Arbuckle, Mauro, and Powell call "the nebulous networks of postmodern knowledge creation and transfer" (2018, para 2). Accessibility, through innovative processes such as versioning, enhances the broad reach of open social scholarship, leading to a democratic engagement across a culturally rich spectrum of participants. But such processes do not necessarily provide coherent critical constellations or knowledge clusters from the perspective of the broad audience, e.g. members of the general public, who may otherwise benefit from the open social methodology. Further, due to the positive benefits of functioning as a group—however geographically

dispersed—open social scholarship teams may ignore or simply not register potential discovery research breakthroughs that do not meet the criteria for the groups' success; in other words, there may be additional breakthroughs that occur in parallel with acknowledged targets met or achieved. In all three instances (knowledge dispersal, lack of knowledge development coherence for *all* of the community and non-community members across a network, and parallel knowledge breakthroughs that remain dispersed/unrecognized), machine learning and topic modelling (see Blei, Ng & Jordan 2003, & Blei 2012) can provide a rigorous methodology for recognizing and understanding open social knowledge creation. Arbuckle, Mauro, and Powell suggest that "large scale analysis" of open data is crucial:

> While the free-to-read argument is important, it is equally important that OA allows for large scale analysis. JSTOR's Data for Research portal is a good start for such access, but a truly OA world would not require restrictions. Full text data mining across many journals and over many years can allow for patterns in research and new directions that are not currently possible because no single individual can read such large collections. HathiTrust's recently released Data API (application programming interface) is another noteworthy step toward a world in which large scale analysis is the norm, thereby increasing the speed and efficiency of research (para 15).

Examples from two different communities are highly relevant and are explored in this paper: The first is Draux and Szomszor's *Topic Modelling of Research in the Arts and Humanities: An analysis of AHRC grant applications* (2017) and the second is Liu and Jansson's "Topic modelling analysis of Instagram data for the Greater Helsinki region" (2017); in both instances, topic modelling can mine publicly funded and/or publicly available relatively "big" data for understanding what would otherwise remain archived and/or dispersed (see, also, Hong & Davison 2010). My approach synthesizes these two approaches to dispersed data—the academic and the commercial—to further theorize topic modelling as a tool for making more transparent the profound benefits of open social scholarship. While many digital humanists have explored topic modelling, either experimenting with large data sets or producing serious research outputs with this tool, my paper also hints at other benefits and issues in adopting—and adapting—this technology for open social scholarship.

Finding Hidden Gems

Topic modelling is not, of course, a homogeneous methodology, and there are many available approaches to this sort of machine reading of big data; in the instance of Draux and Szomszor's report, "Non-negative Matrix Factorisation" (NMF) was considered preferable to the now mainstream LDA (Latent Dirichlet Allocation)/Mallet approach (2017, 15; see Dhillon and Sra 2016), although Greene, in an Insight Report for the Centre for Data Analytics, suggests that one of the challenges of NMF is "instability" or, significantly different results produced "for different runs on the same data matrix" (no date, 10). Draux and Szomszor do allude to the instability challenge near the end of their report, providing a useful and concise explanation of "stability" where "the stability of topics" is defined as the "variation of topic existence when comparing to the two adjacent smaller topic models and the two adjacent bigger topic models" (16). The simple version of the reasoning behind the choice of NMF versus LDA is that NMF downplays frequent words in favour of "infrequently used terms" (Draux and Szomszor, 15; see also, Jones 1972). But what does all of this mean in relation to the actual topic models that are produced? O'-Callaghan, et. al., argue that LDA works well with broad data sets, while NMF is better for drilling-down into specialized material: "While LDA may offer good general descriptions of broader topics, our results indicate that the higher coherence and lower generality associated with NMF topics mean that the latter method is more suitable when analyzing niche or non-mainstream content" (2015, 5656). Yet a warning needs to be sounded here, and that is that *both* LDA and NMF are potentially "unstable" (given the above definition) because of stochastic (randomly occurring or unpredictable) processes that occur in the "initialization phase" of the topic modelling algorithms, i.e., a "random component [that] can affect the final composition of the topics found and the rankings of the terms that describe these topics" (Belford, Namee, & Greene 2018, 159). It often comes as something of a surprise to discover that topics can transform themselves so profoundly during the iterations of the topic modelling algorithm, to such an extent that some of the words that comprise the topic model "may appear or disappear completely between runs" (ibid). As Belford, Namee, and Greene argue in their introduction (and most digital humanists would undoubtedly agree): "...it is clear that any individual run should not be treated as a 'definitive' summary of the underlying topics present in the data" (ibid). The challenge of interpretability of topic models is another relevant and significant issue (see below), with a large literature covering different potential solutions, either at the level of software or methodology/hermeneutics; such a challenge is also raised by Draux and Szomszor in their short concluding section "Interpreting and Labelling the Topics," and is returned to at the end of this paper. But all of the above warnings and cautionary

comments are not meant to suggest that topic modelling does not work, merely that some thought and methodological refinement is needed when thinking about the data sets being mined—i.e., short tweets, metadata, entire essays, chapters or books, and so on.

Analyzing Arts and Humanities Research Council grant applications using topic modelling is a significant step forwards in approaching data that is not usually data mined and is a prime example of a methodology that can work across large dispersed open knowledge data sets. David and Wingrove (Foreword) suggest that this approach "has unlocked rich information within the funding data, not through the existing discipline taxonomies but through analysis of the content of application summaries" (1). In their Introduction, Draux and Szomszor also compare their approach with traditional methodologies that pre-determine in many respects the knowledge domains within which the particular contribution or breakthrough is to be found, e.g., reductive "bibliometrics" or "top-down" models such as library classification systems and journal-based knowledge domains or "fields" (2). In contrast, topic modelling is "a bottom-up approach driven by the material that is available" (ibid):

> Topic modelling is a promising approach that can capture trends in research production. It can map documents to time patterns and spatial distribution. Potentially, combined with expert interpretation, it could leverage information to create insights on emerging and branching topics. Digital science uses this powerful tool to study publications, grants, case studies and other documents and provide understanding of topic evolution, clusters and trends (ibid).

The question of "expert interpretation" is important, even prior to working with the actual topics produced: unsupervised machine learning may sometimes be preferred (e.g., with extremely large data sets), or a hybrid approach can be taken where "controls" are introduced to focus the predictive modelling as such. The latter is known as interactive topic modelling (ITM), "...an in situ method for incorporating human knowledge into topic models" (Hu, Boyd-Grabber, and Satinoff 2011, 248). Two striking phrases from the foreword and introduction of *Topic Modelling of Research in the Arts and Humanities,* as quoted above, are "unlocked rich information" and "capture trends in research production"; the former implies that topic modelling can reveal not just "information" per se, but information that contains valuable insights, whereas the latter implies that topic modelling moves the researcher from a static model of data to one of generative data. In the world of open knowledge production, the dynamic space of data flow necessarily

needs to be understood as generative, i.e., utilizing new paradigms for understanding that we are in the data flow, and that the data flow in and of itself constitutes new worlds. Here I am alluding to Chen and Venkatachalam's theory that big data "should be perceived as a continuous, unstructured and unprocessed dynamics of primitives, rather than as points (snapshots) or summaries (aggregates) of an underlying phenomenon" (2017, 362). Big data in this model is "...the continuous archive of whatever people said, did and even thought" (ibid., 365).

How far does Draux and Szomszor's report, then, live up to the relatively high expectations that one has after reading the foreword and introduction? In short, the paper can be seen as a significant contribution to topic modelling from the perspective of real-world data mining. For example, the AHRC data that is mined includes "successful and *unsuccessful* applications for grants between 2005 and 2016" (3, my emphasis); "unsuccessful" applications are an example of a knowledge domain that may otherwise escape analysis, perceived academically as a sort of Bermuda Triangle of lost research, except that in the reality, unfunded projects often continue, albeit in modified form—e.g., in my own MeTA Digital Humanities lab at Vancouver Island University (VIU), funding from the Canada Foundation for Innovation and the British Columbia Knowledge Development Fund built the lab itself and supported many years of digital humanities research structured by the overarching goals and framework of investigating image-text intersections; an unsuccessful application to SSHRC, however, led to a shift away from the production side of the image-text project, and an overall transformation of the lab with a conceptual/coding and big data/machine learning focus that has been enormously productive in intellectual and practical terms (including using machine learning in the classroom, and for student-focused research projects in the lab). In other words, the project continued regardless of the funding, but in profoundly altered form. How, then, would the public access this new MeTA DH Lab overarching project framework (goals, methodology, even such original components as ideas and aspirations)? It might seem obvious and redundant to note that grant applications contain not only the core intellectual idea(s), but also the methodological framework or structure; it is worth repeating the obvious here because future researchers could utilize both the germ of the idea(s) and the methodology—given an awareness via topic modelling—of the significance of this or any other unfunded application, in other words, through data mining grant applications.

Further practical and concrete advantages of topic modelling in *Topic Modelling of Research in the Arts and Humanities* include tracking "the discovered topics to reveal trends and highlight patterns obscured by the sheer scale of the overall corpus" (Draux and

Szomszor, 9), as well as providing large-scale data visualizations, such as the Grants Similarity Network (9-11) and the Topic Heatmap (12-13). While the former works with similarity clusters and connectivity across networks, the latter works with correlations to reveal "informative patterns": "The heatmap... makes use of additional metadata provided by applicants for each grant to compare a pre-existing categorical structure to the emergent topic model" (12). Commenting on the Grants Similarity Network diagram in their "Outcomes" section, Draux and Szomszor suggest that:

> The large network diagram... shows that such a model can provide an overall landscape of the content that captures many different aspects of the research. This kind of backbone can then be used as a basis for comparison, e.g., by highlighting the different areas of the network that correspond to applications from particular universities. Alternatively, this may uncover interdisciplinary work where clusters of applications appear that link different topics, or it may identify areas of research that are relatively isolated (14).

Such advantages do need to be offered with some reservations in mind, e.g., Draux and Szomszor's suggestion that a "fundamental limitation of topic modelling is that it is unable to parse sentence structure or understand language semantics" (p.14). In many respects, this can be turned into a strength when machine reading large quantities of data, since topic modelling proves itself to be remarkably good at working with reasonably comprehensive units of meta-data. In other words, other digital humanities tools can be utilized to drill down into the "linguistic" data once the topic modelling has been undertaken at the "meta" level. So while it is true that "polysemy can create problems with interpretation because single words are used in different contexts with different meanings" (Draux and Szomszor, 14), in the case of LDA, there are some simple solutions available.

Machine Reading Dispersed Networks

Returning, then, to networked and highly distributed and dispersed open social scholarship, Draux and Szomszor make a relevant point about interpretation: "...topic modelling provides a categorical framework that is driven by the text content alone. It is not determined by pre-existing heuristic beliefs about what the content contains" (14). As noted above, open social scholarship groups may ignore or simply not register potential discovery research breakthroughs that do not meet the criteria for the groups' success; subsequently, there may be additional *unrecognized knowledge breakthroughs* that occur in par-

allel with *recognized or acknowledged* targets met or achieved. I have been using the AHRC report as an exemplary approach to topic modelling data, yet it is important to register that open social scholarship is usually far more dispersed than such a database, i.e., across fluid networks of exchange such as Twitter (see Grandjean 2016), where knowledge production is as likely to take place through conversations and comments on or via social media as any other expressive or representative mode. "Conversations" of course include even shorter or pithier components, such as hashtags, whose symbolic functioning far outweighs the length or complexity of the word or phrase attached to the hashtag. "Symbolic" here means that different metaphors can transform our understanding of how the data linked to the hashtag can spread, e.g., utilizing the metaphor of "infectiousness" and biological-statistical models (see Skaza and Blais 2017). Liu and Jansson (2017) utilize topic modelling to work with such a dispersed domain of data, in this instance, topic modelling Instagram data concerning the Greater Helsinki Region; while this approach most directly serves the needs of tourism initiatives, there are valuable lessons for scholarly enquiry. Examining the benefits of social media data versus traditional methods of feedback on a topic, Liu and Jansson foreground the fact that social media data is "timely, [and] geo-tagged… with fine-grained location data, rich demographics and more context information" (2). A key phrase used by Liu and Jansson for this data is "community knowledge," which, in the context of their project, facilitates "innovative analytical approaches for understanding important issues in urban planning and regional development" (ibid). Topic modelling the "text content extracted from the Posts and Comments fields" of the Instagram data, raises some intriguing methodological questions and observations, such as the importance of hashtags "as effective content" (3; see also 7-8), as well as significant shifts in their analytical results when, for example, removing comments: "…when removing Comments from content input it caused a considerable change to the amount of data as well as the proportion of words and hashtags in the content" (7).

What Liu and Jansson have encountered in this instance, is one of the challenges of working with short texts within conventional LDA, something that has been explored by Jonsson and Stolee (2016), who compare and contrast three topic modelling methodologies: "aggregating documents by author," "the biterm topic model" (BTM), and clustering "word2vec vectors using a Gaussian mixture model" (1). Jonsson and Stolee's experimental paper is interesting because they are in effect testing the results of *the* classic paper in this field: Yan, Guo, Lan, and Cheng's "A biterm topic model for short texts" (2013), which attempts to tackle the shortcomings of LDA, literally speaking, with shorter texts

such as tweets; rather than adopting the usual co-occurrence model between word and document, the "biterm" model examines "unordered word pairs" and models these across the entire corpus (1-2). In replicating and testing Yan, Guo, Lan, and Cheng's research, Jonsson and Stolee find: unexpected results, some minimal differences between topic modelling approaches on short texts (i.e., having analyzed H scores and coherence scores), but overall conclude that "BTM was superior to all other models when working with short documents" (9). Automated procedures for evaluating topic coherence scores were proposed—and proven to be highly effective—by Lau, Newman, and Baldwin in 2014, so even this time-consuming task can now be streamlined.

On Interpretation

Draux and Szomszor raise some hermeneutic issues in the final section of their Report, where they distinguish between themes, categories, and word collocations, since topics "represent the words that appear together in documents, regardless of their meaning" (16). A whole host of interpretive issues therefore conclude the report:

> While they often bring together related terms that align well with concepts such as research discipline, location, methodology or stakeholder group, [the topics] can also reveal idiomatic or pragmatic features of the text corpus. For example, research documents such as grant applications or article abstracts will often contain non-research content such as copyright statements or phrases about the purpose of the research. These will be captured by the topic model, but can be filtered out.
>
> Topics can be labelled for convenience, with the best results achieved using input from domain experts. However, the labelling process can lead to over-interpretation since a human will draw on background experience to infer relationships between terms that may not be present in the text (ibid).

For the purposes of machine reading/data mining large-scale dispersed open social scholarship, I suggest that it is best *not* to filter out the "idiomatic or pragmatic features of the text corpus" as these may, first, flag up important clusters of knowledge breakthroughs or insights (i.e., that do not fit the expected pattern or paradigm), and second, even with copyright statements (to stick to the above example), that there are important lessons to be learnt concerning the access frameworks within which particular academic conversations, projects, and outputs occur. Topic labelling is notorious for being an act of

interpretation that can anchor a particular topic model in a homogenous field of interrelationships, when the fact is that topic models are closer to critical constellations, that is to say, heterogeneous clusters of generative terms that created a particular document or set of documents within a corpus. While labelling is perhaps necessary if topic modelling is being used as a form of automated annotation of big data, in the case of open social scholarship I suggest that the entire topic model *is* the label, that is to say, a discursive field rather than that field represented by a reductive word/concept (see Underwood's (2012) highly perceptive suggestion that topics are "the discourses... that could have generated the documents"). As such, and returning to notions of "instability" and "topic incoherence" from a humanist (and phenomenological) perspective, it is the breakdown of topic modelling (both literally, i.e., breaking down the entire process which leads to understanding its algorithms, and metaphorically, i.e., as a car or an individual can break down) that may indicate or express some generative discourse that would otherwise have been missed. This idea is analogous to Heidegger's suggestion that we only really notice—and understand—the role of equipment in our lives, when that equipment is no longer "ready-to-hand," i.e., it becomes "the un-ready-to-hand" which Heidegger suggests "...can be encountered not only in the sense of that which is unusable or simply missing, but as something un-ready-to-hand which is *not* missing at all and *not* unusable, but which 'stands in the way' of our concern" (103). In other words, incoherent topic models—the prime example of topic modelling idiomatic features—are "un-ready-to-hand" because they are disturbing to us; they foreground not just the fact that LDA, for example, is a tool that sometimes does not function as expected, but that it now "stands in the way" (ibid) of our often fixed, or at the least solidified, interpretive expectations. I suggest a creative application of this analogy: first, coherent topic models are tools or equipment that are "ready-to-hand" and subsequently we don't really notice or focus on them, as they successfully annotate what we expected or wanted to find in the corpus (coherent "themes" that can subsequently be humanly labelled); second, incoherent topic models are "un-ready-to-hand" and in their breakdown they disturb the smooth flow of annotation/mapping via themes, and instead, disclose something unknown or unexpected. As such, incoherent topic models can "permanently" interrupt the smooth flow of automated big data analysis; but as Dreyfus points out in his commentary on *Being and Time*: "Once our work is permanently interrupted, we can either stare helplessly at the remaining objects or take a new detached theoretical stance towards things and try to explain their underlying causal properties" (1995, 79). While this may appear to be stating the obvious to any highly motivated, curious digital humanities researcher, the following sen-

tence is key: “Only when absorbed, ongoing activity is interrupted is there room for such theoretical reflection” (ibid).

The question remains, then, if topic modelling can enable researchers to see past hermeneutic horizons, or, to put it another way, if topic modelling can see beyond the cultural, theoretical, and ideological paradigm(s) through which research takes place? I am hinting here that this may be an important question at a time when universities appear to some academics to have become politically homogeneous in terms of their faculty and non-faculty personnel, as well as in terms of the dominant ideological and theoretical filters through which university members perceive the world (see Langbert, Quain, and Klein 2016; Pew Research Center 2016; Jaschik 2017; Heterodox Academy Mission Statement). Open social scholarship that utilizes social media as a highly positive and dynamic mode of communication also runs the risk of simply replicating such ideological homogeneity and group think in ways analogous to the functioning of social media platforms in general. For example, the desire for more public ownership of scholarly publishing, while understandable when research is publicly funded, does not necessarily make economic sense from a free market perspective (e.g., publicly owned and managed institutions are notorious for having an infinite capacity for subsidized growth). I am not arguing one way or another here, simply suggesting that in an ideologically homogeneous environment, participants in open social scholarship conversations often mention “self-evident” truths on topics such as mechanisms for scholarly publishing, which are *anything but self-evident from other political and intellectual perspectives*. I am advocating utilizing topic modelling precisely *upon* these important and fascinating conversations to get a sense, once more, of what might be called “idiomatic” outcomes/solutions: perhaps a hybrid approach that has been articulated but not recognized, a network of conversations that are so broadly disseminated and dispersed that no conclusions have been reached from them, an opinion that does not “fit” the political climate of the day, and so on. But my argument, it must be said, is not restricted to any particular academic period. I believe that all of the outputs from the arts and humanities that have ever been digitized and will be digitized (or, of course, are born digital), should be part of this data mining process.

References

Arbuckle, Alyssa, Aaron Mauro, & Daniel Powell. 2017. Introduction: “tracing the movement of ideas.” *Social Knowledge Creation in the Humanities: Volume 1*. Edited by Alyssa

Arbuckle, Aaron Mauro, and Daniel Powell. Iter Press. https://ntmrs-skc.itercommunity.org/tracing-movement-ideas/

Belford, Mark, Brian Mac Namee, & Derek Greene. 2018. "Stability of topic modeling via matric factorisation." *Expert Systems With Applications*, 91: 159-169. http://dx.doi.org/10.1016/j.eswa.2017.08.047m

Blei, David M. 2012. "Probabilistic communication models." *Communications of the ACM, 55*: 77-84.

Blei, David M., Andrew Y. NG, & Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research, 3*, 993-1022.

Chen, Shu-heng & Ragupathy Venkatachalam. 2017. "Agent-based modelling as a foundation for big data." *Journal of Economic Methodology*, 24 (4): 1-22. doi.org/10.1080/1350178X.2017.1388964

Dhillon, Inderjit & Suvrit S. Sra. 2005. "Generalized nonnegative matrix approximations with Bregman Divergences." *Proceedings of the 18th International Conference on Neural Information Proc. Systems*, Vancouver, BC, Canada.

Draux, Hélène & Martin Szomszor. 2017. *Topic Modelling of Research in the Arts and Humanities: An analysis of AHRC grant applications*. Arts and Humanities Research Council, London, Digital Research Reports.

Dreyfus, Herbert L. 1995. *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division 1*. Cambridge, Massachusetts & London, England: MIT Press.

Grandjean, Martin. 2016. "A social network analysis of Twitter: Mapping the digital humanities community." *Cogent Arts & Humanities*, 3: 1171458 (14 pp.). http://dx.doi.org/10.1080/23311983.2016.1171458

Greene. Derek. nd. *Matrix Factorization For Topic Models*. Centre for Data Analytics. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.702.4867&rep=rep1&type=pdf

Heidegger, Martin. 1990. *Being and Time*. Trans. John Macquarrie & Edward Robinson. Oxford: Basil Blackwell.

Heterodox Academy. Mission Statement. https://heterodoxacademy.org/about-us/#1546895756166-81f7fb41-2c8a

Hong, Liangjie & Brian D. Davison. 2010. "Empirical study of topic modelling in Twitter." *1st Workshop on Social Media Analytics* (SOMA '10), July 25, n.p.

Hu, Yuening, Jordan Boyd-Graber & Breanna Satinoff. 2011. "Interactive topic modelling." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 248–257, Portland, Oregon, June 19-24.

Jaschik, Scott. 2017. "Professors and Politics: What the Research Says." *Inside Higher*

Ed, February 27. https://www.insidehighered.com/news/2017/02/27/research-confirms-professors-lean-left-questions-assumptions-about-what-means

Jones, Karen Spärke. 1972. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation, 28*(1): 11-21; reprinted *Journal of Documentation, 60* (5) 2004: 493-502.

Jónsson, Elías & Jake Stolee. 2016. "An evaluation of topic modelling techniques for Twitter." Research paper. https://www.cs.toronto.edu/~jstolee/projects/topic.pdf

Langbert, Mitchell, Anthony J. Quain & Daniel B. Klein. "Faculty Voter Registration in Economics, History, Journalism, Law, and Psychology." *Econ Journal* Watch, 13(3): 422-451.

Lau, Jey Han, David Newman & Timothy Baldwin. 2014. "Machine reading tea leaves: Automatically evaluating topic coherence and Topic Model Quality." *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, Gothenburg, Sweden, April 26-30.

Lovejoy, Arthur O. 2001. *The Great Chain of Being: A Study of the History of an Idea*. Cambridge, Massachusetts & London, England: Harvard University Press.

Liu, Shuhua & Patrick Janson. 2017. "Topic modelling analysis of Instagram data for the Greater Helsinki region." *Arcada Working Papers*, pp.1-9.

O'Callaghan, Derek, Derek Greene, Joe Carthy, & Pádraig Cunningham. 2015. "An analysis of the coherence of descriptors in topic modeling." *Expert Systems with Applications*, 42: 5645-5657.

Pew Research Center. April 26, 2016. "A Wider Ideological Gap Between More and Less Educated Adults." https://www.people-press.org/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/

Skaza, Jonathan & Brian Blais. 2017. "Modeling the infectiousness of Twitter hashtags." *Physica A*, 465: 289–296. http://dx.doi.org/10.1016/j.physa.2016.08.038

Underwood, Ted. 2012. "Topic modeling made just simple enough." https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan & Xueqi Cheng. 2013. "A biterm topic model for short texts." *WWW 2013*, May 13–17, 2013, Rio de Janeiro, Brazil.

BIO

Professor Richard J. Lane is the Principal Investigator of the MeTA Digital Humanities Lab at Vancouver Island University, and a past Director of Innovation Island Technology

Association, Nanaimo. Lane is the author or editor/co-editor of sixteen academic books, including Doing Digital Humanities (Routledge 2016, co-ed with Crompton and Siemens), The Big Humanities: Digital Laboratories/Digital Humanities (Routledge 2016) and Malcolm Lowry's Poetics of Space (Ottawa UP 2016, co-ed with Mota) which is available in an open access edition: https://ruor.uottawa.ca/bitstream/10393/35740/1/9780776623412_WEB.pdf

Lane's forthcoming book, co-edited with Crompton and Siemens, is called *Doing More Digital Humanities* (Routledge, 2019/2020).

DOI:

10.21810/pop.2019.008

Citation:

Richard J. Lane, 2019. "Dispersed/Networked Open Social Discovery Research: Applications for Humanistic Machine Learning & Topic Modelling." *Pop! Public. Open. Participatory.* no. 1. https://doi.org/10.21810/pop.2019.008.

Abstract:

One of the benefits of open social scholarship also presents researchers with a challenge: the dispersed nature of the knowledge breakthroughs presented by a diverse network of scholars inside and outside of the academy. Accessibility enhances the broad reach of open social scholarship, leading to a democratic engagement across a culturally rich spectrum of participants. But such processes do not necessarily provide coherent critical constellations or knowledge clusters from the perspective of the broad audience. Further, due to the positive benefits of functioning as a group, open social scholarship groups may ignore or simply not register potential discovery research breakthroughs that do not meet the criteria for the groups' success. In all three instances (knowledge dispersal; lack of knowledge development coherence for all of the community and non-community members across a network; parallel knowledge breakthroughs that remain dispersed/unrecognized), machine learning and topic modelling can provide a methodology for recognizing and understanding open social knowledge creation.

Pop! Public. Open. Participatory is published by the Canadian Institute for Studies in Publishing, located on the unceded territories of the xʷməθkʷəy̓əm (Musqueam), Sḵwx̱wú7mesh (Squamish), and Sel̓íl̓witulh (Tsleil-Waututh) Nations. *Pop!* acknowledges the generous support of Publishing@SFU; the Faculty of Communication, Art, & Technology; the Scholcomm Lab; and the Implementing New Knowledge Environments (INKE) partnership. Read more About Pop! and our Colophon & Credits.

Pop! Public. Open. Participatory
ISSN 2563-6111