Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses

2016 Environ. Res. Lett. 11 093001

(http://iopscience.iop.org/1748-9326/11/9/093001)

Environmental Research Letters

TOPICAL REVIEW

Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses

Emily Grubert[1] **and Anne Siders**

Emmett Interdisciplinary Program in Environment and Resources, 473 Via Ortega, Y2E2 Building, Suite 226, Stanford, CA 94305, USA
[1] Author to whom any correspondence should be addressed.

E-mail: gruberte@stanford.edu **and** siders@stanford.edu

Keywords: text mining, life cycle assessment, adaptive capacity, topic modeling, collocation analysis, digital humanities, computational social science

OPEN ACCESS

RECEIVED
7 August 2015

REVISED
9 August 2016

ACCEPTED FOR PUBLICATION
26 August 2016

PUBLISHED
16 September 2016

Abstract

Digitally-aided reviews of large bodies of text-based information, such as academic literature, are growing in capability but are not yet common in environmental fields. Environmental sciences and studies can benefit from application of digital tools to create comprehensive, replicable, interdisciplinary reviews that provide rapid, up-to-date, and policy-relevant reports of existing work. This work reviews the potential for applications of computational text mining and analysis tools originating in the humanities to environmental science and policy questions. Two process-oriented case studies of digitally-aided environmental literature reviews and meta-analyses illustrate potential benefits and limitations. A medium-sized, medium-resolution review (~8000 journal abstracts and titles) focuses on topic modeling as a rapid way to identify thematic changes over time. A small, high-resolution review (~300 full text journal articles) combines collocation and network analysis with manual coding to synthesize and question empirical field work. We note that even small digitally-aided analyses are close to the upper limit of what can be done manually. Established computational methods developed in humanities disciplines and refined by humanities and social science scholars to interrogate large bodies of textual data are applicable and useful in environmental sciences but have not yet been widely applied. Two case studies provide evidence that digital tools can enhance insight. Two major conclusions emerge. First, digital tools enable scholars to engage large literatures rapidly and, in some cases, more comprehensively than is possible manually. Digital tools can confirm manually identified patterns or identify additional patterns visible only at a large scale. Second, digital tools allow for more replicable and transparent conclusions to be drawn from literature reviews and meta-analyses. The methodological subfields of digital humanities and computational social sciences will likely continue to create innovative tools for analyzing large bodies of text, providing opportunities for interdisciplinary collaboration with the environmental fields.

Background

Emergence of text analysis tools

Scholars have a common and long-standing interest in reviewing large bodies of literature. Indeed, a literature review is a common prerequisite for scholars to demonstrate how their efforts build upon and engage previous work. However, an accelerating rate of publication and growing body of literature renders comprehensive reviews logistically challenging. Many reviews address widely cited, canonical texts from a field, but a critical examination of a field's canon, including its representativeness and how various works came to be included or excluded, relies on an understanding of the full literature. Furthermore, comprehensive analysis may reveal patterns, gaps, and assumptions not apparent in a selective review. Enhanced methods to interrogate large bodies of literature are therefore needed.

Text mining, here defined as the practice of applying computational tools to derive meaning from sets of documents too large to manually review, originated in the humanities. Early examples include computational analysis of the works of Aquinas by Roberto

Busa in the 1940s, of Early Middle High German texts by Roy Wisbey, and of poetry by Stephen Parrish in the 1960s (Hockey 2004). Digital humanities, the disciplinary intersection of computer sciences and humanities, has existed as a subfield since the 1990s, and application of its computational text mining methods is growing more common in humanities scholarship (e.g. Allison *et al* 2013, Katsma 2014, Algee-Hewitt and McGurl 2015, Duhaime 2016, Hoyt *et al* 2014). Similar methods are being applied in the social sciences, particularly in a subfield called computational social science (e.g. Laver *et al* 2003, Cardie and Wilkerson 2008, Lazer *et al* 2009, Yu *et al* 2011), and in medical fields (e.g. Oertelt-Prigione *et al* 2010).

The term text mining encompasses many tools and methods, and many are derived from numerical computational tools developed for use in computer science, biology, and other natural science fields (Blei *et al* 2003). Word frequency analysis, text clustering, sentiment analysis, and topic modeling are used in marketing (Sullivan 2001), security (Corney *et al* 2002, Gegick *et al* 2010), policy (Talamini *et al* 2012), and stakeholder preference analysis in engineering (Castro-Herrera and Cleland-Huang 2010), among others. Text analysis tools are also used in academic research, particularly in the humanities and social sciences where major sources of information and reported research results are text based rather than numerical. Application of machine learning tools can identify and explore constructs and patterns that would otherwise be inaccessible due to the massive amounts of information involved (Grimmer 2015).

Computational tools have been applied in environmental fields (e.g. Kostoff *et al* 2008, Neff and Corley 2009, Altaweel and Bone 2012, Zamagni *et al* 2012, Vasara *et al* 2013, Cody *et al* 2015, 2016, Convertino *et al* 2016) using methods like bibliometrics and media analysis to describe the state of a field and to recommend interpretations and research directions. A growing number of bibliometric analyses, which assess patterns in bibliographic data, demonstrates interest by environmental researchers in use of such tools (e.g. Kugo *et al* 2005, Janssen 2007, Neff and Corley 2009, Altaweel and Bone 2012, Zamagni *et al* 2012, Vasara *et al* 2013, Wang *et al* 2014, Chen *et al* 2016). Authors in other fields have suggested the use of computational methods to aid literature reviews (e.g. Kostoff *et al* 2001, Juola 2008, Ananiadou *et al* 2009), but the utility has not yet been demonstrated in environmental sciences, and use of text mining techniques for environmental analyses remains limited.

In this review, we report on two projects to conduct digitally-aided analyses in climate change-related areas of environmental science: environmental life cycle assessment (LCA) and adaptive capacity to climate change. Our experience illustrates opportunities for text mining and digital tools to enhance environmental science reviews and meta-analyses by enabling large-scale literature reviews, improving replicability and transparency, identifying latent assumptions and gaps within empirical work that are not visible at a manual scale, and lowering barriers for incorporating humanist and social science-type questions in environmental sciences.

Why a systematic review of digital tools in environmental science is needed

Environmental science and policy research is producing increasingly policy-relevant and, in many cases, highly time-sensitive publications related to shifts in climate change and other processes that are 'pressing, pervasive, and uncertain' (Stirling 2006). The authority accorded the Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC) reflects the need for synthesis of environmental science, although the IPCC reports also reflect the contentious and labor-intensive process of manually reviewing thousands of publications (IPCC 2013). Computational methods assist hypothesis testing at scale and allow authors to summarize and describe field-wide trends with more comprehensive data about the literature than is possible with more conventional selective citation and quotation. This access to field-scale information also enhances authors' ability to check their own work for biases or oversights. In addition to making large reviews more feasible, computational methods enable analysis of qualitative data for large-scale patterns, increase replicability and transparency of subjective work, and decrease barriers to interdisciplinary work through shared methods. Meta-analyses are valuable for identifying trends and gaps in literature, building consensus, and generating field-wide research agendas (e.g. Geist and Lambin 2002, Reap *et al* 2008, Aguinis and Glavas 2012, Zamagni *et al* 2012, McManamay *et al* 2013). However, traditional meta-analyses have been limited to fields with quantitative results. Computational text mining provides tools to quantitatively assess qualitative and narrative data, as is increasingly common in work on social-ecological systems. Text mining also increases the ability of meta-analyses to identify trends quickly and at scale by using Moretti's concept of 'distant reading' to identify patterns in large bodies of literature (Moretti 2013). As seen with the benefits of 'big data' analysis in other fields, many patterns of interest may not be visible at a manual scale. While text mining tools cannot naively determine the quality, importance, and validity of a work, they can be extremely useful for researchers applying their expert judgment in the form of testable heuristics (figure 1).

Digitally-aided reviews reduce subjectivity and increase transparency by requiring the researcher to characterize and identify assumptions in review methods, thereby increasing replicability. This is particularly important for multi- and interdisciplinary areas of environmental science, such as climate vulnerability assessments, where colleagues might have different

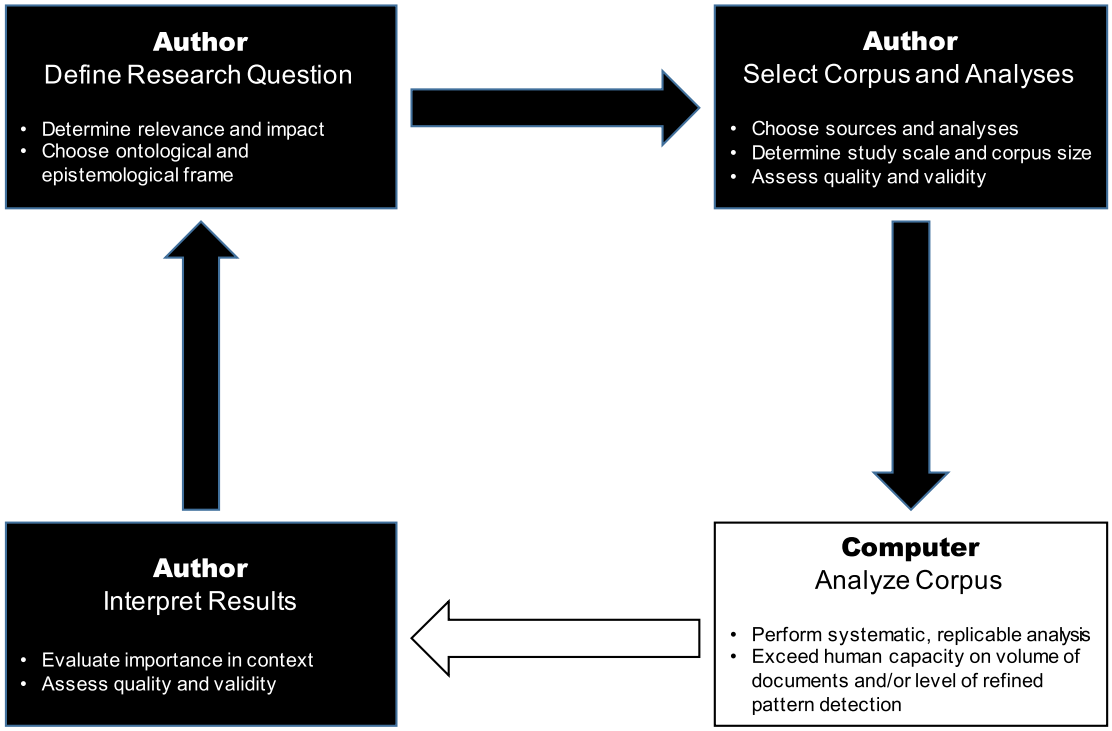

Figure 1. While text mining tools can dramatically increase the speed and consistency of evaluating large volumes of textual data versus a human analyst, human judgment is extremely important to high quality research.

ontological, epistemological, and methodological norms and standards (Moon and Blackman 2014). Selective literature reviews may reflect disciplinary or expertise bias of the reviewer in selecting which works to include, and the standards by which texts are included or excluded may be opaque even to the author. Digital methods, by requiring the researcher to clearly articulate the limits of the analysis to a computer, increase authors' awareness of selection criteria and, ideally, are made transparent to readers. For example, literature reviews are often limited to pieces of both topical relevance and high methodological quality; however, the quality standards for inclusion are rarely articulated. Digital reviews can similarly filter inclusion, but the criteria, based on publication source, timing, or author affiliation, must be stated explicitly. This serves both to inform novice readers about quality standards in the field and to make quality criteria transparent and open for discussion. Further, results of digital analyses applied broadly can be used to help articulate how high quality studies differ from the average.

Digital tools are not a replacement for expert scholars (Yu *et al* 2011), and, indeed, substantial expertise is often needed to develop effective selection criteria and interpret results—especially as it relates to judging the quality and relevance of work. Rather, digital tools enhance experts' ability to gather information and facilitate communication with non-expert audiences by promoting clear explication of subjective choices and by providing a means for non-experts to test expert hypotheses and validate trends that appear at small close-reading scales. Since computational tools can be operated without the background knowledge or theoretical basis an expert might have (Ryan and Bernard 2003), they can be helpful in attracting new perspectives on familiar information. In this way, digital reviews can lower barriers to entry in a field (Hoover 2013) and promote interdisciplinary perspectives. Further, given the global nature of many environmental problems and climate change in particular, methods for evaluating multilingual literature—long considerations in the humanities and linguistics (e.g. Kammer 1989, Sinclair *et al* 1998)—can be of great value, although translingual evaluation remains a challenge.

Finally, applying digital humanities tools to scientific literature analysis presents an unusual combination of methods and ideas between disciplines that do not conventionally adhere to the same general assumptions about the origin and nature of knowledge. In so doing, it has the potential to question underlying assumptions about how scholars acquire and analyze knowledge in their field. The use of humanist and social science digital tools could represent an important opportunity to introduce other forms of humanistic and social scientific inquiry to the environmental sciences, opening the way to systematic critical analysis of possible biases in the environmental science literature. As an already inherently interdisciplinary area of inquiry, the environmental sciences are particularly well suited to the integration of

humanities and social science methods. Given the relevance of environmental science to policy development and its potential to inform decisions about core societal outcomes, particularly related to climate change and other mechanisms of global environmental change, biases within the field can have high stakes. We as a community have a responsibility to investigate these biases. As the community gains facility with digital methods, digitally-aided reviews could advance to the level of novel meta-research, asking questions like how science influences policy; how diversity in the sciences affects the types of questions being addressed through research; or how institutional or disciplinary biases affect results. Being able to review literatures more completely and with a more empirical basis could be an effective way to raise the salience of concerns about representativeness and objectivity (e.g. Harding 1998, Gibson-Wood and Wakefield 2013) while preserving qualitative inquiry for in-depth examination of the mechanisms behind such questions (Yu *et al* 2011).

Text mining tools are increasingly common and particularly well suited to literature investigations at the scale of an entire field or subfield. Many tools do not require extensive computational expertise, and an increasing number of software packages (both general, such as the TM package for R, and specific, such as MALLET for topic modeling) are becoming available. However, exposure to these tools might be unusual for environmental scientists who do not work adjacent to researchers in the digital humanities, computational social sciences, or machine learning. This review highlights important existing work on methods for digitally-aided literature reviews and illustrates the utility of these methods for environmental science through two case studies. Further, it suggests possible areas for application, discusses the implications of this work for environmental science, and notes some limitations and caveats that should inform effective application. This review briefly touches on methods from digital humanities, computational social sciences, and computer science that are likely to be useful to those interested. A much broader bibliography curated for interested scholars can be found on Zotero via DARIAH (2012).

Outcomes and hypotheses

Our main objective is to introduce a suite of computational tools that are increasingly common in the social sciences and humanities to the environmental science and engineering communities by describing the tools, the expected benefits of their use, and procedural experience from two case studies deploying such tools as examples. Thus, outcomes will be greater familiarity with text mining tools and their application to environmental science and, potentially, an increased number of researchers who consider using large-scale computational methods to conduct critical literature reviews and meta-analyses.

Our specific case studies, described in the next section, test the hypotheses that:

- The academic LCA literature reflects a long-term and notable shift in focus toward climate change at the expense of focus on other impacts, particularly human health. The computationally aided review discussed in this article uses topic modeling to demonstrate a clear shift in topics over time.
- The academic literature on climate-relevant adaptive capacity does not use a consistent set of measurable determinants of success, but connections between and among determinants can be used to construct a mechanism-based model for adaptive capacity. The computationally aided review discussed in this article uses network analysis, collocation analysis, and natural language processing to complement manual coding and indicate relationships across determinants as described in the field's literature to date.

Case studies: digitally-aided meta-reviews in the environmental sciences

Case study 1: Topic modeling the LCA literature

The first case study presented in this work focuses on the use of topic modeling in the LCA literature. It is intended to illustrate the suggestion that digital tools can aid hypothesis-driven literature reviews at a scale that cannot be performed manually: this case study investigates the titles and abstracts of about 8200 journal articles, essentially the complete English language LCA corpus published in the peer reviewed literature between 1995 and 2014 (Grubert 2016). This case study also illustrates the value of digital tools for enabling replicable testing of hypotheses developed from a more traditional reading program. After reading about 300 LCA method and practice articles, the author perceived that treatment of climate issues was becoming more pronounced in the more recent literature while treatment of more traditional pollution was becoming rarer. However, given a literature of thousands of articles, making a definitive claim that this trend existed was uncomfortable. Topic modeling provided a way to test this hypothesis at a scale beyond typical reading capacities.

Case study 2: Exploring and validating connections among qualitative concepts of adaptive capacity

The second case study is a meta-analysis of literature on the adaptive capacity of social systems to climate change. Rather than expand the size of the body of literature reviewed, this study used digital tools to perform in-depth analyses on a relatively small data set of 275 full-length academic articles and non-academic reports, comprising 88% of all academic work in the field (Siders, in prep). The use of text analysis tools

permitted recovery and analysis of information at a micro-scale not visible to human reader comprehension. Manual coding was used to identify the field's first comprehensive list of 165 determinants of adaptive capacity used in the field to date. Collocate analysis and network visualization were then used to assess how determinant terms relate to one another within texts. Results raise questions about assumptions and theories currently held, consciously or unconsciously, by researchers in the field and inform the development of a new mechanism-based model of adaptive capacity. The study illustrates how an iterative process of computational tools and manual review permits meta-analysis of qualitative and narrative data to test hypotheses, identify assumptions, and generate new research directions and theory at a level not previously accessible.

Methods

We present two case studies of the use of text mining tools for literature review and analysis in the environmental sciences in this article, focusing on process and research development rather than results. More results-oriented descriptions can be found in Grubert (2016) and Siders (in prep).

Both case studies represent 'large' datasets from a traditional review and meta-analysis perspective, analyzing 8239 and 275 texts respectively. The key difference in the two case studies is in their level of analysis. The first uses computational analysis to approximate human interpretation of the texts by identifying macro-level trends through topic modeling. Notably, this macro-view relies on word frequency rather than order, enabling the use of bag-of-words data. The second focuses on word order and syntagmatic analysis (what words occur near one another) to recover information at micro-scales invisible to human reading comprehension. In both analyses, the computational tools permit analyses that would not be possible by human readers, in the first because the body of texts is too large and in the second because the patterns are too small. The studies are selected in this manner to highlight the diversity of approaches that can be undertaken with computational text analysis tools.

Article selection criteria

Case study 1: Life cycle assessment, Web of Knowledge

Over 8200 articles were selected based on a May 2014 Web of Knowledge search for entries with topic = 'life cycle' (in quotation marks), publication date between January 1995 and May 2014, and language = English. Non-English language articles were excluded from this analysis because topic modeling is not well suited to a corpus with untranslated inclusions, as the non-majority language is often grouped into a single topic by the model regardless of meaning. Fewer than 2% of articles discovered using the 'life cycle' topic were not in English, but this is likely related to the use of an English language search term. Filters were applied based on author judgment and manual checks to exclude most articles dealing with concepts like the life cycle of an organism rather than one of the methods associated with LCA principles. For example, the categories 'ecology' and 'psychiatry' were excluded due to large volumes of research on biological and drug life cycles. Categories included were: environmental sciences, engineering environmental, energy fuels, engineering civil, engineering manufacturing, management, engineering industrial, operations research management science, business, environmental studies, public environmental occupational health, water resources, forestry, planning development, sociology, and urban studies, further refined to the research areas of environmental sciences ecology, energy fuels, public environmental occupational health, sociology, social sciences other topics, and mining mineral processing. Results were restricted to entries classified as articles or reviews due to the focus of this study on trends in the academic journal literature. Of the resulting 8239 articles selected, 2107 (26%) were categorized as United States-based, while 6132 (74%) were not United States-based. This result is consistent with the idea that English is a common language for journal articles even in regions and countries that are not primarily English-speaking. The full dataset, classified by year of publication, is available online (Grubert 2016). Note that Web of Knowledge has been somewhat reorganized since May 2014; as of 2016, the closest equivalent search group would be Web of Science, 'all databases.'

Case study 2: Adaptive capacity, Web of Knowledge

A January 2016 Web of Knowledge search for entries in all databases with title = 'adaptive capacity' (in quotation marks) published between 1800 and 2015 yielded 529 results, 448 of which were non-duplicate English language academic articles. Based on reading titles, journal titles, and abstracts where necessary, author judgment was used to classify results as relating to adaptation of a social system or a non-social system. For example, articles relating to the adaptive capacity of neurons, spines, pigs, and rats were considered 'non-social.' Social response includes response of businesses and organizations, response of communities to climate change or environmental shifts, and response of social-ecological systems. Where an article was in question, it was included in the corpus. Non-academic reports, such as handbooks, guidebooks, and analyses by non-profit organizations, were included if they were under 50 pages, to avoid the length of the longer reports from biasing frequency analyses. Eight long reports were eliminated. Full-length, English language texts were needed for the collocate analysis, so 14 non-English texts were excluded. The relative scarcity of non-English texts indicates English is the dominant language of this field.

This filtering yielded 295 articles and reports, of which 261 (88%) full-length texts could be located in open-access or Stanford library resources. As this is an emerging field that frequently references a small set of articles from related fields of vulnerability assessment and hazards geography, which may not use the term 'adaptive capacity' in the title, an additional search for entries with topic = 'adaptive capacity' was run and sorted based on citation frequency. Of the top 20 most cited entries, 14 were related to social capacity, and these were added to the corpus for a final collection of 275 articles. Full-length texts were converted to .txt files for analysis, and title information and references were removed to focus analysis on the text body.

Digital review methods

Case study 1: Topic modeling

Case study 1 uses topic modeling to assess trends in the LCA literature over time. Topic modeling is a computational technique that associates individual words in a document to a cluster of words called a topic, creating a many-to-many mapping between documents and a collection of k topics. The value of k is a user choice, often driven at least partially by the study goal. For example, k might be over 100 if the goal is to find unusual topics in a highly dispersed corpus, while it might be less than 10 for a study like this one with the goal of identifying common themes in a tight corpus (i.e., one that is stylistically self-similar, in this case because all the documents are journal abstracts that tend to be governed by relatively strict rules).

In addition to determining the number of topics k, a topic modeler chooses from among several computational techniques for allocating words to topics. This study uses a common technique called latent Dirichlet allocation (LDA) using Gibbs sampling, as it was developed for applications involving journal abstracts quite similar to that of this work (Blei *et al* 2003, Blei and Lafferty 2006, Blei and Lafferty 2007). LDA is a mixture model based on the assumption that individual words in a document are there because they are part of a topic addressed by the document, and one of its benefits relative to something like a mixture of unigrams model is its improved ability to distinguish between homonyms based on context (Riddell 2012, Underwood 2012). (Note that a 'document' can be defined to suit the application, ranging from single sentences to collections of writings by a particular author or from a particular university, for example.)

Topics are defined by an algorithm based on how words appear in the overall corpus, and specifically based on how often they appear with other specific words in documents within that corpus. Here, determining word distribution uses Gibbs sampling, which fits a model assuming individual words in documents are exchangeable (that is, the bag-of-words model, which discards word order but preserves word frequency and association to documents). Words are evenly divided across k topics at random, then iteratively reallocated based on two probabilities: the probability that the word appears in a given existing topic and the probability that words in a given document belong to the topic to which the word is currently assigned. Based on these probabilities, a word is either moved to a different topic or retained in its currently assigned topic. When this sampling process has been repeated sufficiently often that randomly selected words are no longer being reallocated, the topics are considered stable. For an intuitive description of the method, see Jockers (2011).

Once topics are defined, words are expected to occur in a document with some probability based on the presence or absence of that topic. Examining which words actually do occur in a document enables inferences about which topics are present. Very common words that appear in almost every topic (like 'and,' 'for,' 'the,' etc), called 'stop words,' are usually removed when the goal is to uncover content rather than stylistic topics (see e.g. Jockers and Witten 2010), as is the case here. Notably, topic modeling produces word lists, not labeled topics: that is, a 'topic' might include 'water river lake lacustrine ecosystem' with high probability, and it is a human task to assign that topic a parsable label. Thus, there remains a significant amount of subjectivity in turning computer-generated topics (word lists) into parsable topic themes (human-generated labels), including the decision of whether and where to truncate word lists for analysis. Technically, all topics include all words in the corpus, but with decreasing posterior probabilities: labeling typically proceeds based on the top n words. By publishing lists of the highest probability words in each topic, scholars can provide evidence that they have chosen a reasonable topic label and invite others to comment. Good overviews of topic modeling as it is applied in practice, including some introduction to techniques other than simple LDA, can be found at e.g. (Riddell 2012, Underwood 2012, Weingart 2012); a topic modeling bibliography is curated by David Minmo at mimno.infosci.cornell.edu/topics.html.

For this case study on LCA in the English language journal literature, the corpus of 8239 English language journal abstracts and titles described above was topic modeled using LDA in MAchine Learning for LanguagE Toolkit (MALLET) (McCallum 2002) via the Topic Modeling Tool UI. Documents were defined as the titles and abstracts of articles published in a given year between 1995 and 2014 (so, a document would be all titles and abstracts for 1996), addressing the question of how topical focus has changed over time.

MALLET was run using 1000 iterations, a topic threshold of 0.05, top 20 word returns, and an augmented stopwords list defined as the *Journal of Machine Learning Research* list plus context-specific additions: 'Elsevier,' 'life,' 'cycle,' 'assessment,' 'lca,' 'rights,' 'reserved,' 'paper,' 'study,' and 'results.'

Sensitivity to the number of topics k was tested by evaluating topics for $k = [2, 10]$; $k = 2, 8, 9$, and 10 were considered nonadditive due to under- and over-differentiation, so results address only $k = [3, 7]$. No stemming, or consolidation of similar words based on roots, was used, as stemmers would consolidate words like 'transport' (as in transport of contaminants) and 'transportation' (as in systems for moving people and products), 'gas' (as in greenhouse gas or natural gas) and 'gasoline' (as in oil-derived fuel), and others that have very different meanings in environmental science.

Topics were hand-labeled after examining top words by topic for MALLET output across multiple runs for each k to ensure robustness. Since words are randomly assigned to k topics at the start of a run, results will not be identical across runs unless the same seed is used. Though the corpus used here is too large to code manually, the study was based on experience from reading and hand-annotating over 300 environment and energy-related LCA articles. More detail on the method and its implementation for this case study can be found in Grubert (2016).

Case study 2: Collocation and network visualization

Case study 2 uses a nontraditional collocation analysis and network visualization to create a mechanism-based model of adaptive capacity. In linguistics, collocation refers to a 'set phrase', or a sequence of words that repeatedly co-occur, or co-locate (Firth 1957, Halliday 2002, Sinclair 1991). Computational linguistics often analyses collocations to explore language learning and development. This study, however, uses a broader concept of collocations in which two concept terms that are co-located with statistically significant frequency are assumed to have a conceptual relationship; the basis for this assumption is that authors use two words in close proximity when expressing a connection between the two. Collocates may also indicate a functional relationship, such as that between a participle and gerund; these were excluded in this analysis but may be present in similar approaches.

Computational collocation analysis calculates the probability that words *a, b, c*, etc will occur within n words of base word x (Halliday 2002). The value of n is a user choice, often selected to correspond roughly to a standard unit of English language, such as immediate pairing, sentence, or paragraph. In this study, a value of $n = 10$ was chosen to approximate the two words appearing within the same sentence (an average English language sentence being 15–20 words in length). The analysis can be symmetric (looking 10 words left and 10 right) or asymmetric (looking 10 words left and 0 right). This study used symmetric analysis. For example, collocate analysis takes the word 'education' and identifies all words that appear within 10 words on either side. It then calculates a measure of association based on how frequently the words appear in the corpus as a whole and determines which pairings are statistically significant. This study uses a Fisher's exact test with a p-value of 0.01 to test for significance, due to the relatively small number of observed co-locations. 'Education' may appear frequently near words 'then' or 'an', but as these are so frequent within the English language, the co-occurrence is unlikely to be significant (or of interest).

Results may be filtered to display only relationships among words of interest, as was done in this study. In this case, a list of terms related to the determinants of adaptive capacity, as identified by a close reading of the texts and review of initial collocation results, was used to filter so that only collocates of two determinant words were displayed in the results. For example, 'education' co-occurring near 'field' was not of interest, as field is not considered a determinant of adaptive capacity. However, 'education' collocating near 'knowledge,' 'wealth,' or 'occupation' was of interest. The determinant word list included 491 terms. As this analysis used full terms, rather than stemmed words (using only the root), terms such as 'resource' and 'resources' or 'resilient' and 'resilience' were listed as separate entries in the determinant word list. Stemming the words before analysis would have removed this need, but it would have also removed the ability to distinguish between words that share a stem but have different conceptual meanings, such as 'adaptive' and 'adaptation,' which were of interest to this analysis. The determinant list included some two-word phrases, such as 'adaptive capacity' and 'human capital.' The collocation analysis located all such two-word phrases in the corpus and joined them before running the collocate analysis (effectively searching for the word 'adaptive_capacity'), so terms such as 'adaptive' were not double-counted both when appearing by itself and as part of the phrase. Although collocation analysis creates a purely statistical association, there is a high degree of subjectivity in selection of the terms of interest to explore within the collocate results. But, by pairing collocate analysis with close reading of the text and word frequency analysis, authors can provide evidence that they have chosen a reasonable set of focus words.

This study is based on the premise that two determinant terms, when used within a sentence-length of one another, are conceptually related. For example, phrases such as 'The stock of human capital including education and personal security' and 'it is important to note that health care and education have strong positive correlations with per capita income' (Yohe and Tol 2002) suggest that 'education' is related to 'human capital,' 'security,' and 'income'. However, on occasion, terms are co-located because they appear within a list, table, or other non-language construct such that their co-occurrence is not indicative of a direct relationship. The relationship between 'psychological' and 'financial resources,' for example, is due to their co-occurrence within several literature reviews that list

these as two determinant factors. Drawing out 'Key Words in Context', an author can quickly examine a word of interest and the sentences immediately surrounding it to manually check the type of relationship being expressed. This was done with a random set of results and with any results that appeared unexpected based on a manual reading of the texts.

Collocate relationships were then visualized as a network of relationships by turning collocate words into nodes and connections (ratio of number of observed collocations to expected collocations) into edges. The 491 determinant terms were collapsed into 165 determinant concepts, using expert judgment. Words that connected to 'resource' were also assumed to connect conceptually to 'resources', so that 'resource' and 'resources' become a single node. Other terms were more subjective, and raise questions for the field as to whether 'values', 'norms', 'attitudes,' and 'beliefs,' for example, should be considered the same or distinct determinants of adaptive capacity. The final network was visualized in Gephi 0.8.2 as a directed network with a Forced Atlas layout.

Visualizing the network allows for visual inspection and interrogation by the author. Several analyses on the network can also be conducted, and these are well-described in the social network analysis literature. In this study, centrality measure and modularity analysis were of particular importance. A betweenness centrality measure is an indicator of the node's relative importance to the network, as it indicates the number of paths between two other nodes that must pass through this central node (Freeman 1977) (Gephi uses an algorithm by Brandes (2001) to measure betweenness). Modularity analysis can also be conducted to measure how well the network separates into modular communities, or groups (Girvan and Newman 2002) (Gephi uses an algorithm by Blondel *et al* (2008) for community detection). More detail on the method and its implementation for this case study can be found in Siders (in prep).

Review results

In both case studies, digital tools were indispensable for testing hypotheses and identifying patterns in environmental science subfields. This conclusion motivates the present work. Results of the reviews described above are presented briefly here and in more detail in Grubert (2016) and Siders (in prep).

Case study 1: Topic modeling identifies trends in the LCA literature

Topic modeling using simple LDA with Gibbs sampling effectively revealed topical shifts over time in the English language LCA literature. Topic proportions and most common words for LCA abstracts and titles by publication year are further discussed in Grubert (2016) and are also available interactively online at lcatopics.emilygrubert.org. Topic modeling provides empirical support for the hypothesis based on manual annotation that climate change has become a far more important topic in LCA over time. Topic modeling adds to this manual intuition by revealing an important secondary narrative: climate change appears to have become more prevalent at the direct expense of attention to human health. This tradeoff trend persists for all coherent k, $k = [3, 7]$.

Case study 2: Collocation analysis and network visualization identifies assumptions and generates research directions

Collocation analysis and network visualization and analysis with measures of centrality and modularity detection enabled exploration of how determinant terms relate to one another. Relationships among determinant terms point toward conceptual relationships and toward possible functional roles of various determinants, as discussed further in Siders (in prep.). This insight into functional roles provides policy-relevant information on how interventions may alter the adaptive capacity of social systems. However, digital analysis does not provide conclusive evidence that the functional roles identified in the literature to date are, in fact, the state of the world; rather, by identifying and quantifying patterns in the literature too small and too complex for manual detection, the digital analysis assesses results of the field to date, identifies assumptions, and generates new questions and directions for research.

Discussion

The environmental sciences can benefit from the use of digital tools for conducting literature and meta-reviews. Beyond simply applying existing tools, the cross-disciplinary work provides an opportunity for deeply inter- and transdisciplinary exchange among natural sciences, engineering, social sciences, and humanities. This discussion section briefly highlights available tools, current limitations, and recommendations for research.

Text mining tools available to the environmental sciences

The ability to perform large-scale, replicable literature reviews and analyses is critical to maximize the benefits of interdisciplinary work on environmental issues. While this review presents case studies using specific text mining tools, computational text mining tools are diverse. The Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) project describes a classification system for digital research based on research activities, objects, and techniques (Borek *et al* 2016). The goal of the taxonomy is to recognize that scholars using text analysis tools might be more united by method than by topic or field, so

Table 1. Common text mining analytical activities and techniques.

Research activity	Common related techniques	Example questions
Content analysis		What do the documents mean, using words as data points?
	Topic modeling	*What are the major themes in the documents?*
	Sentiment analysis	*Do the documents reflect a positive feeling about a topic?*
Network analysis		How are the actors in the documents (e.g. authors, characters, cited authors) related to each other?
	Citation Network Analysis	*How do concepts and terms spread through the field?*
Relational analysis	*Most frequent word analysis Most distinguished word analysis*	How are the documents related to each other, e.g. what is the difference between version 1 and version 2?
Spatial analysis		What spatial patterns do the documents exhibit, e.g. is one country the most frequent object of study?
Structural analysis		Are there common linguistic forms within the documents?
	Collocation analysis	*How do terms used relate to one another conceptually?*
	Most distinguished word analysis	*Which terms are hallmarks of certain subcategories of research?*
	Most frequent word analysis	*Is one form of a multi-word phrase more common than others?*
Stylistic analysis		Can documents be correctly distinguished as originating with a certain author or culture based on the words used?
	Principal component analysis (PCA)	*Are certain words or phrases good differentiators between groups of documents?*
	Cluster analysis	*Are there clear groups of documents in this corpus, e.g. journal articles versus books?*
Visualization		Can major patterns be communicated better in graphical form?

Source: Based on the TaDiRAH framework.

classification improves communication about methodological development and novel applications (Borek *et al* 2016). As an example, a selected summary of analytical activities and techniques expected to be useful in the environmental sciences, alongside examples of the types of research questions they address, is found in table 1. TaDiRAH also maintains classifications for upstream and downstream activities like capture (gathering texts to generate a corpus) and dissemination.

Current limitations of text mining for environmental science

The potential for digital tools to create value in literature reviews and analyses is high, specifically in environmental fields, but prospective users should be aware of limitations. Some are relatively clear, such as the idea that digital tools should not be viewed as a replacement of expert judgment. As with other data generation and analysis tools, digital textual analysis tools can streamline collection, improve ease of analysis, and provide a path to replication of datasets —but they cannot be used to synthesize information, assess quality, draw conclusions, or pose questions independent of human scholarship. This review highlights several possibly nonobvious limitations, the first three of which are related to the academic cultural context of environmental science and engineering.

First, text analysis requires textual data, and since the application of text analysis tools is not yet widespread in the environmental fields, many issues of copyright protection and access have not yet been addressed. For environmental scholars seeking to do analysis based on bag-of-words models where word counts but not word order are available by document, for example those based on word frequency, some prominent environmental journals are not yet included in publisher initiatives to release downloadable bag-of-words data. Publishers typically do not permit large-scale download of full texts (this practice led to a very high profile case when Aaron Swartz of MIT attempted such a download, JSTOR 2011), making acquisition of large full-text corpora labor intensive. However, an increasing number of publishers, including Elsevier, JSTOR, Nature Publishing Group (NPG), Springer and others (Crossref; see tdmsupport.crossref.org for ongoing updates), have committed to policies for the express purpose of enabling certain types of digital analysis. This trend is in response to substantial effort by scholars interested in text analysis (Poole 2013) and is greatly aided by the efforts of university libraries to include full-text and analytic access requirements in contracts with publishers and holders of digital archives. As of 2014, available datasets were not representative of environmental literatures; however, this condition appears to be changing, and increased awareness of the utility of text mining to environmental science could accelerate this change.

A second academic culture point concerns modes of publication and choice of text format. Scholars conducting text analysis retain significant control over the choice of which texts and text formats to include in their analysis, and authors should consider whether the most familiar source for texts in their field is the appropriate one. For example, in the natural sciences, journals are likely the best source of academic information. Engineers and applied natural scientists also commonly publish in conference proceedings and trade journals as a primary mode. Social scientists are more likely to publish in multiple modes, including journals (though often with different journals and even different publishers than natural scientists or engineers), white papers (particularly common for economists), and books (particularly common for anthropologists and sociologists). Humanists commonly publish in monographs and book chapters in addition to articles. Furthermore, digital humanists and computational social scientists might publish in 'born digital' formats that are not yet easily indexed by more traditional search tools, though efforts to give such projects higher visibility and legitimacy are underway (Karampelas 2015). As readers might notice, some of the references for this article are blog posts and other website sources that are not published in traditional academic outlets: the choice to reference these posts rather than exclusively journal articles is an intentional decision meant to highlight the location of much of the most vibrant academic discourse about text analysis. For scholars in fields that do not traditionally reward methodological development or have not yet embraced digital techniques, there are few formal outlets for tool discussion. Thus, the blog world, Github, Twitter, and other online-only sources are increasingly important loci for academic discourse. All these sources might be readily available to a scholar conducting text analysis, but they may not necessarily be accessed through the same search engines and techniques. Investigating whether the chosen corpus is actually the most appropriate one, and whether some branches of relevant knowledge have been excluded by a narrow search, is wise.

Related to the question of corpus inclusiveness is a note that environmental scientists in particular should be sensitive to the fact that there is a history of natural scientists dismissing or devaluing humanist and social scientific inquiry. Given this opportunity to learn from what is now decades of scholarship on developing tools for large scale textual analysis in the humanities, natural scientists should be cautious of a common underlying prejudice by positivists that interpretivist inquiry is less valid and that, since algorithms and computational work have traditionally been the purview of natural science and engineering, natural scientists can apply these tools immediately and without consideration of more interpretivist or constructivist lines of inquiry. There are well developed methods for asking and addressing questions about cultural lenses, themes, networks, and many other issues relevant to the environmental fields in texts, and digital analysis without accompanying expert interpretation can lead to conclusions experts might have interpreted much differently or the omission of key terms that are generally recognized. A proposed solution is to remain respectful of other modes of inquiry, actively seek out scholarship from other fields about the types of questions one is asking (noting the need to search in different modes than might be comfortable noted above), and in some cases, pursue collaborators.

In addition to these cultural challenges, there are also technological challenges to text analysis. Sentiment analysis is a tool of common interest to researchers looking to uncover information about author assumptions (e.g. 'this technology is good'), public opinion (e.g. 'this impact is scary'), policy emphasis (e.g. 'this policy favors biomass'), or similar questions related to emotional valence. While sentiment analysis is an area of improving research (see e.g. Cody *et al* 2015), it retains several major weaknesses when used off the shelf. One is that common sentiment lexicons like the Harvard General Inquirer (Stone *et al* 1962; now at http://www.wjh.harvard.edu/~inquirer/) represent enormous amounts of work but usually require some context-specific modification before they can be applied. For example, whether the word 'cheap' reflects positive or negative sentiment is highly specific to the application. Another weakness is that sentiment analysis tools can struggle with multiword negatives: in a bag of words model where word order is lost, it is difficult to distinguish between 'good' and 'not good.' Similarly, choosing the most appropriate text mining techniques for a given question, identifying cases where an apparent result is an artifact of the method (as with Case Study 2's finding that some collocation of terms is due to authors listing findings of prior studies), and evaluating the relative importance of outputs (as with the global sensitivity and uncertainty analysis performed in Convertino *et al* 2016) remain challenging. Overcoming these technical challenges and modifying text analysis tools appropriately for application in an environmental science context is one more reason to seek collaboration with experts from digital humanities and social sciences.

The final caution offered here is that use of computational tools does not always increase transparency and replicability: software-based analysis can obscure what work was actually completed and thus decrease replicability when authors do not make code available, properly document their steps and particular softwares used, or otherwise reveal their methods (Marwick 2015). Moreover, digital techniques are not without subjective bias: numerous decisions must be made by the author that can introduce bias or affect results. For example, the case studies discussed above included subjective choices about the horizon for collocate analysis, selecting determinants to include or

topics to address, cutting off number of words per topic, and numerous other points. The value of digital techniques lies in making these decisions explicit and transparent, but these benefits are only realized if the authors are open about the choices made. For literature reviews in particular, the risk posed by non-transparent methods might seem slight compared with the opacity of the existing process, but the power of digital analysis to produce results quickly and easily —even absent proper interpretation—underscores the importance of reporting methods clearly, accurately, and completely.

Recommendations for the application of digital text mining tools in environmental fields

Digitally aided literature review and meta-analysis techniques hold substantial promise for improving the accessibility and empiricism of literature-based conclusions in the environmental sciences and beyond. This review ends by highlighting several potential classes of application that could be fruitful.

Text analysis permits scholars to investigate a wide range of questions ranging from field description—what topics does this literature cover? are authors using the same keywords and definitions for major concepts in this literature?—to more intricate questions aimed at revealing underlying directionality in research. As indicated by our case studies, techniques like topic modeling can help to characterize a field and potentially identify gaps and trends in the literature. Similarly, network analysis can be used by a scholar who is reasonably certain that the entire relevant literature is captured to show where poorly connected scholarly communities might benefit from further integration. Word frequency analysis can help identify effective keywords and help scholars understand not only what words they should be searching for but also what words they should use in their own work to maintain coherence in a multidisciplinary field. Other exciting areas of inquiry could investigate whether scholars from particular backgrounds or particular universities ask different questions or treat results differently. Style analysis can also help answer questions like whether major grant calls affect the topics scientists choose to work on or whether certain topics are written more like government versus industry documents. All of these questions can foster further inquiry into questions of why and whether these things matter.

The environmental fields are already transdisciplinary and transepistemological, drawing scholars of the environmental humanities, environmental justice, life sciences, earth sciences, engineering, and many others towards topics that are central to many aspects of human life. Addressing environmental challenges requires effective collaboration across very different perspectives, both in society and in academia specifically. Digital tools for literature investigation can provide direct advantages to someone doing work in the environmental fields, including the ability to review large bodies of literature, to empirically examine hypotheses about the content of existing work, and to identify opportunities for contributions. As important, though, might be the opportunity to engage in truly and deeply transdisciplinary collaboration that sharing a method across fields with dramatically different modes of inquiry could bring.

Acknowledgments

Thank you to Mark Algee-Hewitt for guidance on the case studies presented here and for very helpful feedback on this article. This material is partly based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-114747 and by a David and Lucille Packard Stanford Graduate Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Stanford University or the National Science Foundation.

References

Aguinis H and Glavas A 2012 What we know and don't know about corporate social responsibility a review and research agenda *J. Manage.* **38** 932–68

Algee-Hewitt M and McGurl M 2015 *Between Canon and Corpus: Six Perspectives on 20th-Century Novels* Stanford University Literary Lab Pamphlet 8 (https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf)

Allison S, Gemma M, Heuser R, Moretti F, Tevel A and Yamboliev I 2013 *Style at the Scale of the Sentence* Stanford University Literary Lab Pamphlet 5 (http://litlab.stanford.edu/LiteraryLabPamphlet5.pdf)

Altaweel M and Bone C 2012 Applying content analysis for investigating the reporting of water issues *Comput. Environ. Urban Syst. Spec. Issue: Adv. Geocomput.* **36** 599–613

Ananiadou S, Rea B, Okazaki N, Procter R and Thomas J 2009 Supporting systematic reviews using text mining *Soc. Sci. Comput. Rev.* **27** 509–23

Blei D M and Lafferty J D 2006 Correlated topic models *Adv. Neural Inf. Process. Syst.* **18** 147 (http://galton.uchicago.edu/~lafferty/pdf/ctm.pdf)

Blei D M and Lafferty J D 2007 A correlated topic model of science *Ann. Appl. Stat.* **1** 17–35

Blei D M, Ng A Y and Jordan M I 2003 Latent Dirichlet allocation *J. Mach. Learn. Res.* **3** 993–1022 (www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)

Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E 2008 Fast unfolding of communities in large networks *J. Stat. Mech.* **10** 1000

Borek L, Dombrowski Q, Perkins J and Schöch C 2016 TaDiRAH: a case study in pragmatic classification *Digit. Humanit. Q.* **10** (http://digitalhumanities.org/dhq/vol/10/1/000235/000235.html)

Brandes U 2001 A faster algorithm for betweenness centrality *J. Math. Sociol.* **25** 163–77

Cardie C and Wilkerson J 2008 Text annotation for political science research *J. Inf. Technol. Politics* **5** 1–6

Castro-Herrera C and Cleland-Huang J 2010 Machine Learning Approach for Identifying Expert Stakeholders *2nd Int. Workshop on Managing Requirements Knowledge (MaRK'09)* (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5457348)

Chen H, Jiang W, Yang Y, Yang Y and Man X 2016 State of the art on food waste research: a bibliometrics study from 1997 to 2014 *J. Cleaner Prod.* (doi:10.1016/j.jclepro.2015.11.085)

Cody E, Reagan A, Mitchell L, Dodds P and Danforth C 2015 Climate change sentiment on twitter: an unsolicited public opinion poll *PLos One* **10** e0136092

Cody E, Stephens J, Bagrow J, Dodds P and Danforth C 2016 Transitions in climate and energy discourse between hurricanes katrina and sandy *J. Environ. Stud. Sci.* (doi:10.1007/s13412-016-0391-8)

Convertino M, Munoz-Carpena R and Murcia C 2016 'Reading the minds' for quantitative sustainability: assessing stakeholder mental models via probabilistic text analysis *Information, Models, and Sustainability* ed J Zhang *et al* (Cham: Springer) pp 21–38

Corney M, de Vel O, Anderson A and Mohay G 2002 Gender-preferential text mining of e-mail discourse *Proc. 18th Annual Computer Security Applied Conf.*(doi:10.1109/CSAC.2002.1176299)

Crossref Text and Data Mining (http://tdmsupport.crossref.org/) (Accessed: 3 March 2016)

DARIAH 2012 Doing Digital Humanities—A DARIAH Bibliography (https://zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography)

Duhaime D E 2016 Textual reuse in the eighteenth century: mining eliza haywood's quotations *Digital Humanit. Q.* **10** (http://digitalhumanities.org/dhq/vol/10/1/000229/000229.html)

Elsevier Text and Data Mining (https://elsevier.com/about/company-information/policies/text-and-data-mining) (Accessed: 3 March 2016)

Firth J R 1957 *Papers in Linguistics* (Oxford: Oxford University Press) pp 1934–51

Freeman L C 1977 A set of measures of centrality based on betweenness *Sociometry* **40** 35–41

Gegick M, Rotella P and Xie T 2010 Identifying security bug reports via text mining: an industrial case study *7th IEEE Working Conf. on Mining Software Repositories* (doi:10.1109/MSR.2010.5463340)

Geist H J and Lambin E F 2002 Proximate causes and underlying driving forces of tropical deforestation *BioScience* **52** 143–50

Gibson-Wood H and Wakefield S 2013 'Participation', white privilege and environmental justice: understanding environmentalism among hispanics in Toronto *Antipode* **45** 641–62

Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci.* **99** 7821–6

Grimmer J 2015 We're all social scientists now: how big data, machine learning, and causal inference work together *PS: Polit. Sci. Polit.* **48** 80–3

Grubert E 2016 Implicit prioritization in life cycle assessment: text mining and detecting metapatterns in the literature *Int. J. Life Cycle Assess.* (doi:10.1007/s11367-016-1153-2)

Halliday M A K 2002 Categories of the theory of grammar, 1961, in *Word*, 17(3) *On Grammar: The Collected Works of M.A.K. Halliday* vol 1 (London: Continuum)

Harding S G 1998 *Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies* (Bloomington, IN: Indiana University Press)

Hockey S 2004 The history of humanities computing *Companion to Digital Humanities (Blackwell Companions to Literature and Culture)* ed S Schreibman, R Siemens and J Unsworth (Oxford: Blackwell) (http://www.digitalhumanities.org/companion/)

Hoover D L 2013 Textual analysis *Literary Studies in the Digital Age* ed K M Price and R Siemens (New York: Modern Language Association of America) (http://dlsanthology.commons.mla.org/textual-analysis/)

Hoyt E, Ponto K and Roy C 2014 Visualizing and analyzing the hollywood screenplay with scriptthreads *Digital Humanit. Q.* **8** (http://digitalhumanities.org/dhq/vol/8/4/000190/000190.html)

Intergovernmental Panel on Climate Change (IPCC) 2013 IPCC Factsheet: What Literature Does the IPCC Assess? (https://ipcc.ch/news_and_events/docs/factsheets/FS_ipcc_assess.pdf)

Janssen M A 2007 An update on the scholarly networks on resilience, vulnerability, and adaptation within the human dimensions of global environmental change *Ecol. Soc.* **12** 9

Jockers M 2011 The LDA buffet is now open; or, latent dirichlet allocation for english majors Matthew L. Jockers (http://matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/) (Accessed: 2 March 2016)

Jockers M and Witten D M 2010 A comparative study of machine learning methods for authorship attribution *Lit. Linguist. Comput.* **25** 215–23

JSTOR 2011 JSTOR Statement: Misuse Incident and Criminal Case | About JSTOR (http://about.jstor.org/news/jstor-statement-misuse-incident-and-criminal-case) (Accessed: 3 March 2016)

JSTOR DFR: About Data For Research (http://dfr.jstor.org/??view=text& &helpview=about_dfr) (Accessed: 3 March 2016)

Juola P 2008 Killer applications in digital humanities *Lit. Linguist. Comput.* **23** 73–83

Kammer M 1989 Wordcruncher: problems of multilingual usage *Lit. Linguist. Comput.* **4** 135–40

Karampelas G 2015 Stanford University Press Awarded $1.2 Million for the Publishing of Interactive Scholarly Works | Stanford University Libraries *Stanford University* (http://library.stanford.edu/news/2015/01/stanford-university-press-awarded-12-million-publishing-interactive-scholarly-works)

Katsma H 2014 *Loudness in the Novel* Stanford University Literary Lab Pamphlet 7 (http://litlab.stanford.edu/LiteraryLabPamphlet7.pdf)

Kostoff R N, Block J A, Solka J L, Briggs M B, Rushenberg R L, Stump J A, Johnson D, Lyons T J and Wyatt J R 2008 Literature-related discovery (LRD): lessons learned, and future research directions *Technol. Forecast. Soc. Change* **75** 276–99

Kostoff R N, Toothman D R, Eberhart H J and Humenik J A 2001 Text mining using database tomography and bibliometrics: a review *Technol. Forecast. Soc. Change* **68** 223–53

Kugo A, Yoshikawa H, Shimoda H and Wakabayashi Y 2005 Text mining analysis of public comments regarding high-level radioactive waste disposal *J. Nucl. Sci. Technol.* **42** 755–67

Laver M, Benoit K and Garry J 2003 Extracting policy positions from political texts using words as data *Am. Political Sci. Rev.* **97** 311–31

Lazer D *et al* 2009 Computational social science *Science* **323** 721–3

Marwick B 2015 How computers broke science—and what we can do to fix it *The Conversation* (http://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938) (Accessed: 3 March 2016)

McCallum A K 2002 MALLET: A Machine Learning for Language Toolkit (http://mallet.cs.umass.edu) (Accessed: 26 May 2015)

McManamay R A, Orth D J, Kauffman J and Davis M M 2013 A database and meta-analysis of ecological responses to stream flow in the South Atlantic Region *Southeastern Naturalist* **12** 1–36 (www.eaglehill.us/SENAonline/articles/SENA-mon-5/03-McManamay.shtml)

Moon K and Blackman D 2014 A guide to understanding social science research for natural scientists *Conservation Biol.* **28** 1167–77

Moretti F 2013 *Distant Reading* (London: Verso) (http://amazon.de/Distant-Reading-Franco-Moretti/dp/1781680841)

NPG Nature Publishing Group Participating in Copyright Clearance Center's New Text Mining Solution—Copyright Clearance Center (https://copyright.com/nature-publishing-group-participating-in-cccs-new-text-mining-solution/) (Accessed: 3 March 2016)

Neff M W and Corley E A 2009 35 years and 160 000 articles: a bibliometric exploration of the evolution of ecology *Scientometrics* **80** 657–82

Oertelt-Prigione S, Parol R, Krohn S, Preißner R and Regitz-Zagrosek V 2010 Analysis of sex and gender-specific research reveals a common increase in publications and marked differences between disciplines *BMC Med.* **8** 70

Poole A H 2013 Now is the future now? The urgency of digital curation in the digital humanities *Digit. Humanit. Q.* **7** (http://digitalhumanities.org/dhq/vol/7/2/000163/000163.html)

Reap J, Roman F, Duncan S and Bras B 2008 A survey of unresolved problems in life cycle assessment *Int. J. Life Cycle Assess.* **13** 374-88

Riddell A 2012 A Simple Topic Model (Mixture of Unigrams) (https://ariddell.org/simple-topic-model.html)

Ryan G W and Bernard H R 2003 Techniques to identify themes *Field Methods* **15** 85–109

Sinclair J 1991 *Corpus, Concordance, Collocation* (Oxford: Oxford University Press)

Sinclair J, Mason O, Ball J and Barnbrook G 1998 Language independent statistical software for corpus exploration *Comput. Humanit.* **31** 229–55

Stirling A 2006 Analysis, participation and power: justification and closure in participatory multi-criteria analysis *Land Use Policy* **23** 95–107

Stone P J, Bales R F, Namenwirth J Z and Ogilvie D M 1962 The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information *Behav. Sci.* **7** 484–98

Sullivan D 2001 *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales* (New York: Wiley)

Talamini E, Caldarelli C E, Wubben E F M and Dewes H 2012 The composition and impact of stakeholders' agendas on us ethanol production *Energy Policy* **50** 647–58

Underwood T 2012 Topic modeling made just simple enough *The Stone and the Shell* (http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/)

Vasara P, Rouhiainen J and Lehtinen H 2013 Resource convergence and resource power: towards new concepts for material efficiency *Phil. Trans R. Soc.* A **371** 20110562

Wang B, Pan S-Y, Ke R-Y, Wang K and Wei Y-M 2014 An overview of climate change vulnerability: a bibliometric analysis based on web of science database *Nat. Hazards* **74** 1649–66

Weingart S 2012 Topic modeling for humanists: a guided tour *The Scottbot Irregular* (http://scottbot.net/HIAL/?p=19113)

Yohe G and Tol R S J 2002 Indicators for social and economic coping capacity—moving toward a working definition of adaptive capacity *Glob. Environ. Change* **12** 25–40

Yu C H, Jannasch-Pennell A and DiGangi S 2011 Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability *Qualitative Rep.* **16** 730–44 (http://nsuworks.nova.edu/tqr/vol16/iss3/6/)

Zamagni A, Amerighi O and Buttol P 2012 Finding life cycle assessment research direction with the aid of meta-analysis *J. Ind. Ecol.* **16** S39–52