2017 International Conference on Data Science and Advanced Analytics

Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation

Shaheen Syed
Department of Information and Computer Sciences
Utrecht University
Utrecht, The Netherlands
Email: s.a.s.syed@uu.nl

Marco Spruit
Department of Information and Computer Sciences
Utrecht University
Utrecht, The Netherlands
Email: m.r.spruit@uu.nl

***Abstract*—This paper assesses topic coherence and human topic ranking of uncovered latent topics from scientific publications when utilizing the topic model latent Dirichlet allocation (LDA) on abstract and full-text data. The coherence of a topic, used as a proxy for topic quality, is based on the distributional hypothesis that states that words with similar meaning tend to co-occur within a similar context. Although LDA has gained much attention from machine-learning researchers, most notably with its adaptations and extensions, little is known about the effects of different types of textual data on generated topics. Our research is the first to explore these practical effects and shows that document frequency, document word length, and vocabulary size have mixed practical effects on topic coherence and human topic ranking of LDA topics. We furthermore show that large document collections are less affected by incorrect or noise terms being part of the topic-word distributions, causing topics to be more coherent and ranked higher. Differences between abstract and full-text data are more apparent within small document collections, with differences as large as 90% high-quality topics for full-text data, compared to 50% high-quality topics for abstract data.**

I. Introduction

There is an ever-growing amount of scientific literature with which scientists must grapple and which threatens to overwhelm their capacity to stay up to date with new research [1]. As a consequence, increased availability of tools and algorithms is necessary to match the ever-growing scientific output [2]. These tools and algorithms could aid in exploring large document collections in alternative and structured new ways in contrast to traditional searches. This is especially important as the topics within articles, the main ideas within articles that can be shared among similar articles, cannot always be detected through traditional keyword searches [3].

Topic models are machine-learning algorithms to uncover hidden or latent thematic structures (i.e. topics) from large collections of documents [4]–[7]. The latent thematic structures automatically emerge from the statistical properties of the documents and, as such, no prior labeling or annotation is necessary. In turn, the thematic structures can be used to automatically categorize or summarize documents up to a scale that would be impossible to do manually. Topic modeling approaches have proved to be very helpful in elucidating the key ideas within a set of documents [8]–[10], and they do so with greater speed and a quantitative rigor that would otherwise be possible only through a traditional narrative review [9].

One of the most popular and highly researched topic models is latent Dirichlet allocation (LDA) [6]. LDA is a generative probabilistic topic model that overcomes the limitations of other well-known topic model algorithms such as Latent Semantic Indexing (LSI) [4] and probabilistic Latent Semantic Indexing (pLSI) [5]. LDA models documents as multinomial distributions over K latent topics and each topic is modeled as a multinomial distribution over the fixed vocabulary V. As such, LDA captures the heterogeneity of research topics or ideas within scientific publications and can be viewed as a mixed membership model [11].

Utilizing LDA to uncover latent topics from textual data has been successfully applied in several research domains. Griffiths and Steyvers [8] performed LDA on 28,154 abstracts of the journal *Proceedings of the National Academy of Sciences* (PNAS) to uncover topics and to illustrate their relation to the journal's categorization scheme. Gatti *et. al.* [12] used LDA on 80,757 abstracts from 37 primary journals from the fields of operations research and management science (OR/MS) to gain insight into the historical and current publication trends. A similar approach was performed within the field of transportation research on 17,163 abstracts from 22 leading transportation journals [13] and within the field of conservation science on 9,834 abstracts [14]. Besides being performed on abstract data, LDA has also been applied to 12,500 full-text research articles within the field of computational linguistics [15], 2,326 articles from Neural Information Processing Systems papers (NIPS) [16], and 1,060 articles within agricultural and resource economics [17].

However, the reason for choosing abstract data over full-text data, or vice versa, when using LDA has not been argued for. Although some researchers (e.g. [12]) mention that abstract data is likely to contain a high density of words, thus making it suitable for LDA, others simply mention the dataset without

978-1-5090-5004-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DSAA.2017.61

explaining the rationale for the choice. There are various reasons why this might be the case: one might simply only have access to abstract data (i.e. availability), one may want to keep the computational time to a minimum (i.e. feasibility), or one may want to reduce the pre-processing steps that are often necessary when dealing with full-text articles (i.e. simplicity). These pre-processing steps could include scraping the publishers' repositories; converting PDF to plain-text, either direct or with the aid of optical character recognition (OCR) software; or an increased boilerplate cleaning phase. However, a more scientific rationale is required to aid in the choice of abstract or full-text data when uncovering latent topics with LDA.

This research is the first to explore the practical effects of choosing abstract or full-text data when uncovering latent topics with LDA. In particular, it shows the practical effects when revealing latent semantic structures from documents concerning scientific research publications. The differences between topics are calculated with a topic coherence measure [18]–[21] that shows, in contrast to the likelihood of held-out data, a higher correlation with human topic ranking data, the gold standard for topic interpretability. The underlying idea of topic coherence is rooted in the distributional hypothesis of linguistics [22]—namely, words with similar meanings tend to occur in similar contexts. Additionally, we use the knowledge of a domain expert to rank topics, thus providing, along with topic coherence, a comparison of topic quality from a human perspective.

II. Background

A. Latent Dirichlet Allocation

LDA is a generative probabilistic topic model that aims to uncover latent or hidden thematic structures from a corpus *D*. The latent thematic structure, expressed as topics and topic proportions per document, is represented by hidden variables that LDA posits onto the corpus. The generative nature of LDA describes an imaginary random process based on probabilistic sampling rules from which we assume that the documents come from. However, we only observe the words within documents and need to infer the hidden structure, that is, the topics and topic proportions per document, by applying statistical inference techniques. This process aims to answer the question: Which hidden structure or topic model is most likely to have generated these documents? In doing so, we obtain the posterior distribution that captures the hidden structure given the observed documents. The generative process is defined as follows:

1) For every topic $k = \{1, ..., K\}$
   a) draw a distribution over the vocabulary *V*, $\beta_k \sim \text{Dir}(\eta)$
2) For every document *d*
   a) draw a distribution over topics, $\theta_d \sim \text{Dir}(\alpha)$ (i.e. per-document topic proportion)
   b) for each word *w* within document *d*
      i) draw a topic assignment, $z_{d,n} \sim \text{Mult}(\theta_d)$, where $z_{d,n} \in \{1, ..., K\}$ (i.e. per-word topic assignment)
      ii) draw a word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$, where $w_{d,n} \in \{1, ..., V\}$

Each topic β_k is a multinomial distribution over the vocabulary V and comes from a Dirichlet distribution $\beta_k \sim \text{Dir}(\eta)$. Additionally, every document is represented as a distribution over K topics and come from a Dirichlet distribution $\theta_d \sim \text{Dir}(\alpha)$. The Dirichlet parameter α denotes the smoothing of topics within documents, and η denotes the smoothing of words within topics. The joint distribution of all the hidden variables β_K (topics), θ_D (per-document topic proportions), z_D (word topic assignments), and observed variables w_D (words in documents) is expressed by (1):

$$p(\beta_K, \theta_D, z_D, w_D|\alpha,\eta) = \prod_{k=1}^{K} p(\beta_K|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k}) \tag{1}$$

Fig. 1 shows the LDA probabilistic graphical model in plate notation [23], where the unshaded nodes represent the hidden random variables, the shaded nodes the observed random variables, and the edges the conditional dependencies between them. The rectangles, called plates, represent replication. The graphical model is equivalent to the joint probability of all the hidden and observed variables expressed in (1). We have K topics β_K (K-plate) as distributions over words depending on the Dirichlet parameter η, i.e. $\prod_{k=1}^{K} p(\beta_K|\eta)$. For all D documents (D-plate) we have a per-document topic proportion θ_d depending on the Dirichlet parameter α, i.e. $\prod_{d=1}^{D} p(\theta_d|\alpha)$. Finally, for all N words (N-plate) of a document $d \in D$, we find that the per-word topic assignment $z_{d,n}$ depends on the previously drawn per-document topic proportion θ_d, and the drawn word $w_{d,n}$ depends on the per-word topic assignment $z_{d,n}$ and all the topics $\beta_{d,k}$, i.e. $\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k}$ [we retrieve the probability of $w_{d,n}$ (row) from $z_{d,n}$ (column) within the $K \times V$ topic matrix].

The per-word topic assignment, the per-document topic distribution, and the topics are the latent variables and are not observed. We would have to condition on the only observed variable, the words within the documents, to infer the hidden structure with statistical inference. This can be viewed as a reversal of the generative process. The conditional probability, also known as the posterior, is expressed by (2):

$$p(\beta_K, \theta_D, z_D|w_D) = \frac{p(\beta_K, \theta_D, z_D, w_D)}{p(w_D)} \tag{2}$$

Unfortunately, computation of the posterior is intractable due to the denominator [6]. The marginal probability $p(w_D)$

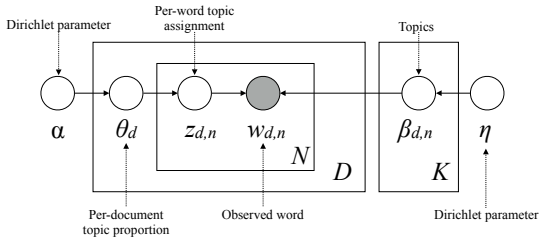


Fig. 1. LDA represented as a graphical model in which the nodes denote the random variables and the edges the dependencies between them. Unshaded nodes are unobserved or hidden variables and the shaded nodes represent the observed random variables. The boxes, called plates, indicate replication.

is the sum of the joint distribution over all instantiations of the hidden structure and is exponentially large [24]. Although the posterior cannot be computed exactly, a close enough approximation to the true posterior can be achieved with statistical posterior inference. Mainly two types of inference techniques can be discerned: sampling-based algorithms (e.g. [25], [26]) and variational-based algorithms (e.g. [27]–[29]). It is important to note that both variational and sampling-based algorithms provide similarly accurate results [30].

B. Topic Coherence Measurement

After approximating LDA's posterior distribution, the K topics are represented as multinomial distributions over V. Each topic distribution contains every word but assigns a different probability to each of the words. The words within topics with high probability are words that tend to co-occur more frequently. These high-probability words, usually the top 10 or top 15, are used to interpret and semantically label the topics. However, LDA outputs as many topics as are defined by K: a low K results in too few or very broad topics, whereas a high K results in uninterpretable topics or topics that ideally should have been merged. Choosing the right value of K is thus an important task in topic modeling algorithms, including LDA.

Measures such as the predictive likelihood of held-out data [31] have been proposed to evaluate the quality of generated topics. However, such a measure correlates negatively with human interpretability [32], making topics with high predictive likelihood less coherent from a human perspective. This is especially important when generated topics are used for browsing document collections by users or understanding the trends and development within a particular research field. As a result, researchers have proposed topic coherence measures, which are a qualitative approach to automatically uncover the coherence of a topic [18], [33], and the underlying idea is rooted in the *distributional hypothesis* of linguistics [22]; words with similar meanings tend to occur in similar contexts. The topics are considered to be coherent if all or most of the words, for example, the topic's top N words, are related. The computational challenge is to obtain a measure that correlates highly with human topic ranking data, such as topic ranking data obtained by word and topic intrusion tests [32]. Human topic ranking data are often considered to be the gold standard, and consequently a measure that correlates well is a good indicator of topic interpretability. A recent study by Röder *et. al.* [19] systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data. Additionally, new coherence measures obtained by combining existing elementary elements were explored as well. Their systematic approach revealed a new unexplored coherence measure, which they labeled C_V, to achieve the highest correlation with all available human topic ranking data. As a result, this study adopts the C_V coherence measure for topic coherence calculations. C_V is based on four parts: (i) segmentation of the data into word pairs, (ii) calculation of word or word pair probabilities, (iii) calculation of a confirmation measure that quantifies how strongly a word set supports another word set, and finally (iv) aggregation of individual confirmation measures into an overall coherence score.

(i) Data segmentation pairs each of the topic's top-N words with every other top-N word. Let W be the set of a topic's top-N most probable words $W = \{W_1, ..., W_N\}$, S_i a segmented pair of each word $W' \in W$ paired with all other words $W^* \in W$, and S the set of all pairs defined as $S = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\}$. For example, if $W = \{w_1, w_2, w_3\}$, then a pair $S_i = (W' = w_1), (W^* = w_1, w_2, w_3)$. Such segmentation measures the extent to which the subset W^* supports, or conversely undermines, the subset W' [34].

(ii) Probabilities of single words $p(w_i)$ or the joint probability of two words $p(w_i, w_j)$ can be estimated by Boolean document calculation, that is, the number of documents in which (w_i) or (w_i, w_j) occurs, divided by the total number of documents. The Boolean document calculation, however, ignores the frequencies and distances of words. C_V incorporates a Boolean sliding window calculation in which a new virtual document is created for every window of size s when sliding over the document at a rate of one word token per step. For example, document d_1 with words w results in virtual documents $d'_1 = \{w_1, ..., w_s\}$ and $d'_2 = \{w_2, ..., w_{s+1}\}$, and so on. The probabilities $p(w_i)$ and $p(w_i, w_j)$ are subsequently calculated from the total number of virtual documents. In contrast to Boolean document calculation, the Boolean sliding window calculation tries to capture the word token proximity to some degree.

(iii) For every $S_i = (W', W^*)$, we calculate a confirmation measure ϕ that calculates how strongly W^* supports W' and is based on the similarity of W' and W^* in relation to all the words in W. To calculate this similarity, W' and W^* are represented as context vectors [18] as a means to capture the semantic support of all the words in W. These vectors $\vec{v}(W')$ and $\vec{v}(W^*)$ are created by pairing them to all words in W, as exemplified in (3). The agreement between individual words w_i and w_j is calculated via normalized pointwise mutual information (NPMI), as shown in (4). NPMI, in contrast to pointwise mutual information (PMI), shows a higher correlation with human topic ranking data [35].

TABLE I
OVERVIEW OF THE DS_1 AND DS_1 DATASETS WHERE J = NUM. JOURNALS; Y = TIME RANGE; D = NUM. DOCUMENTS; N_d = MEAN DOCUMENT LENGTH; N = NUM. TOKENS; V = VOCABULARY SIZE.

	DS_1		DS_2	
	Abstract	Full-text	Abstract	Full-text
J	1	1	12	12
Y	1996-2016	1996-2016	2000-2016	2000-2016
D	4,417	4,417	15,004	15,004
N_d	108.94	3,855.36	123.7	3,850.78
N	481,168	17,029,133	1,856,700	57,777,025
V	14,643	142,852	25,781	379,116

Additionally, ε is used to account for the logarithm of zero and γ to place more weight on higher NPMI values. The confirmation measure ϕ of a pair S_i is obtained by calculating the cosine vector similarity of all context vectors $\phi_{S_i}(\vec{u}, \vec{w})$ within S_i, with $\vec{v}(W') \in \vec{u}$ and $\vec{v}(W^*) \in \vec{w}$ as expressed in (5).

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1,\ldots,|W|} \tag{3}$$

$$\text{NPMI}(w_i, w_j)^\gamma = \left(\frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i,w_j)+\epsilon)} \right)^\gamma \tag{4}$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{5}$$

(iv) The final coherence score is the arithmetic mean of all confirmation measures ϕ.

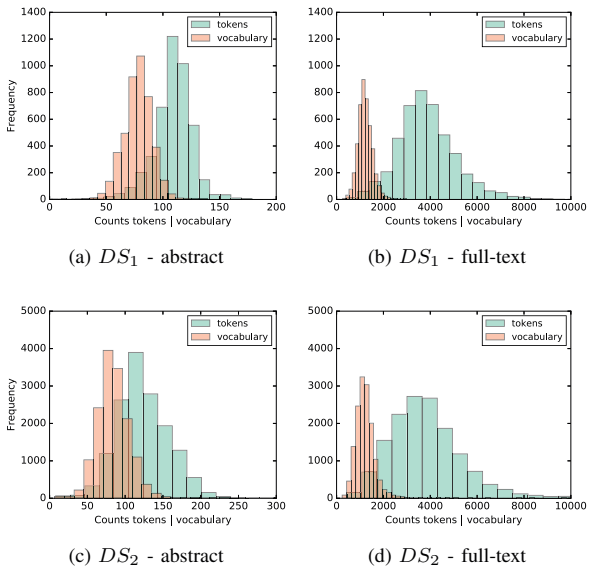
III. METHODOLOGY

A. The Experiment

This paper explores the effects of uncovered latent topics and their topic coherence score, a proxy for topic quality when applying LDA on abstract and full-text data. Besides topic coherence, we explore the effects of human topic ranking—often considered the gold standard for topic interpretability—on topics uncovered from abstract and full-text data. In doing so, we explore the practical effects that types of documents, and more specifically, word length, vocabulary size, and document frequency have on the coherence and interpretability of LDA topics.

B. Dataset

Two datasets were created that contain abstract and full-text data: DS_1 contains 4,417 research articles (1996 to 2016) from the journal *Canadian Journal of Fisheries and Aquatic Sciences*, and DS_2 contains 15,004 research articles (2000 to 2016) from 12 top-tier fisheries journals: *Canadian Journal of Fisheries and Aquatic Sciences*, *Fish and Fisheries*, *Fisheries*, *Fisheries Management and Ecology*, *Fisheries Oceanography*, *Fisheries Research*, *Fishery Bulletin*, *Marine and Coastal Fisheries*, *North American Journal of Fisheries Management*, *Reviews in Fish Biology and Fisheries*, *Reviews in Fisheries Science*, and *Transactions of the American Fisheries Society*. Note that $DS_1 \subset DS_2$ for Y = 2000 to 2016. Regular expressions were used to extract abstracts from full-text articles, as the downloaded articles appeared in full-text.



(a) DS_1 - abstract (b) DS_1 - full-text

(c) DS_2 - abstract (d) DS_2 - full-text

Fig. 2. Histograms of token and vocabulary frequencies for DS_1 and DS_2 for both abstract and full-text data. (b) and (d) contain very long tails for the number of tokens (up to 18,000).

The DS_1 dataset relates to studies where a single scientific journal was analyzed from a domain-specific journal (e.g. [16]), and DS_2 to studies where LDA was used to uncover topics from a multitude of related domain-specific journals (e.g. [12], [13], [15]). The two datasets allow for comparison of not only abstract and full-text data but also on corpus size (i.e. the number of scientific publications). An overview of DS_1 and DS_2 is given in Table I, and histograms of token and vocabulary (i.e. distinct words) frequencies are displayed in Fig. 2.

The choice of these journals was based on two factors: (i) they are domain-specific journals but employ a broad scope of research topics from the field of fisheries, and (ii) a fisheries domain expert was available to manually label and rank the topics as an alternative means of assessing the quality of topics. Furthermore, a domain-specific journal might increase the generalizability to other domain-specific journals (e.g. journals in the domain of social psychology or resource economics) compared to a more general or broadly oriented journal such as *Nature*, *Science*, or *PLOS ONE*. The domain of fisheries includes a multitude of knowledge production approaches, from mono- to transdisciplinary. Biologists, oceanographers, mathematicians, computer scientists, anthropologists, sociologists, political scientists, economists, and researchers from

many other more disciplines contribute to the body of knowledge of fisheries, together with non-academic participants such as decision makers and stakeholders. Within the domain of fisheries, research into text analytics techniques has only been applied in a number of cases (e.g. [36], [37]).

All research articles were downloaded from the journals' repository and converted from PDF to plain text. Full-text data and abstract data were tokenized, and single-character words, numbers, and punctuation marks were removed. Furthermore, we removed all single-occurrence words, words that occurred in more than 90% of the documents, and words that belonged to a standard English stop word list ($n = 153$). Apart from grouping lowercase and uppercase words, no normalization method such as stemming or lemmatization was applied to reduce inflectional and derivational forms of words to a common base form; stemming algorithms can be overly aggressive and could result in unrecognizable words that reduce the interpretability when labeling the topics. Stemming might also lead to another problem, namely that it cannot be deduced whether a stemmed word comes from a verb or a noun [38].

C. Creating LDA Models

For both datasets, and for both abstract and full-text data, we created 40 different LDA models by varying the K parameter (i.e. the number of topics) from 1 to 40 and repeating this process three times (4×120 LDA models in total). The Dirichlet parameters are set to be symmetrical for the smoothing of words within topics $\eta = \frac{1}{V}$ and topics within documents $\alpha = \frac{1}{K}$. By keeping $\alpha < 1$, the modes of the Dirichlet distribution are close to the corners, thus favoring just a few topics for every document and leaving the larger part of topic proportions very close to zero. The LDA models are created using the Python library *Gensim* [39]. Gensim uses variational inference called online LDA [40] to approximate the posterior. The convergence iteration parameter for the expectation step (i.e. E-step) is set to 100; the part where per-document parameters are fit for the variational distributions [see Algorithm 2 in [40]].

D. Topic Coherence

For every LDA model created (480 in total), we calculated the C_V coherence score as explained in Section II-B. Segmentation of top pairs is obtained by pairing every word from the top 15 words with every other word from the top 15 words. In some cases, coherence calculations are based on the top 10 most probable words. However, as no stemming or lemmatization was applied, several words with the same base form were among the top 10 words (e.g. *sample*, *sampling*), so analyzing the top 10 words would effectively mean analyzing less than 10 distinct words. To avoid logarithms of zero when calculating coherence scores, ε is set to a very small number, 10^{-12}, as proposed by Stevens *et. al.* [20]. We furthermore set $\gamma = 1$ to place equal weights on all NPMI values as researched by Röder *et. al.* [19] and have shown the highest correlation with all topic ranking data, in contrast to Aletras and Stevenson [18], where $\gamma = 2$ shows better results. To capture word proximity when calculating word or word pair probabilities, the Boolean sliding window for Boolean document calculation is set to $s = 110$ [19].

The LDA model with the optimal coherence score, obtained with an elbow method (the point with maximum absolute second derivative), was additionally analyzed by a fisheries domain expert. The domain expert is affiliated with the leading competence institution for fishery and aquaculture in Norway. The analysis consisted of an inspection of the top 15 most probable words for each topic, together with an inspection of the document titles and content. Additionally, the domain expert rated the topics (high, medium, low) by assessing the coherence of the top 15 words and the presence of incorrect terms (i.e. words) within each topic. High-quality topics contain no incorrect terms, medium-quality topics contain one or two, and low-quality topics contain three or more. An incorrect term is defined as a word that has no semantic relationship with the topic's top 15 words. The domain expert attached a label to each topic that best captured the semantics of the top 15 words.

IV. Results

Fig. 3 shows the obtained C_V coherence scores for all 480 LDA models created, with Fig. 3a and Fig. 3b displaying the results for the DS_1 and DS_2 datasets, respectively. The lines represent the mean coherence scores from 3 runs where the number of topics was varied from 1 to 40. A visual inspection of Fig. 3a shows that LDA models created with full-text data from the DS_1 dataset achieve higher mean coherence scores among all values of K, a result that is not visible for DS_2 (Fig. 3b).

Table II displays the actual coherence score values for uncovered topics from abstract and full-text data for both datasets. It shows the mean C_V coherence score ($\bar{X}$), the standard deviation (s), and the difference between mean values ($\bar{X}_2 - \bar{X}_1$) calculated from all three runs for $K = \{2, ..., 40\}$. Positive differences between mean values indicate a higher achieved coherence score for full-text data. We furthermore calculate the significance ($p < 0.05$, $p < 0.01$, and $p < 0.001$) between $\bar{X}_1$ and $\bar{X}_2$ with an independent two-sample t-test as Levene's test for homoscedasticity assumes equal variances for all K values.

A. DS_1 Dataset

Although every $\bar{X}_2$ (full-text) outperforms $\bar{X}_1$ (abstract), not all differences are statistically significant. For $k = 3$ and $k = 6$, the differences between coherence scores are not significant but are still very close to the 5% significance threshold. The largest difference between mean values is achieved at $k = 2$ (two-topic LDA model), although it is only significant at $p < 0.05$. Looking at all K values, three runs achieve $p < 0.05$ significance, 18 achieve $p < 0.01$ significance, and 16 achieve $p < 0.001$ significance. The choice of full-text data results in overall topics with higher coherence for all K values, and these differences are significant for all but two LDA models. The abstract data achieved the optimal coherence score (via

TABLE II

CALCULATED COHERENCE SCORE FOR ABSTRACT AND FULL-TEXT DATA FOR BOTH DATASETS. $\bar{X}$ = MEAN COHERENCE SCORE, s = STANDARD DEVIATION COHERENCE SCORE, $\bar{X}_2 - \bar{X}_1$ = DIFFERENCE IN MEAN COHERENCE SCORES, t = CALCULATED T-STATISTIC, p = TWO-TAILED P-VALUE, K = NUMBER OF TOPICS

	Dataset DS_1 (4,417 documents)							Dataset DS_2 (15,004 documents)						
	Abstract$_1$		Full-text$_2$		Statistics (t-test)			Abstract$_1$		Full-text$_2$		Statistics (t-test)		
K	$\bar{X}_1$	s_1	$\bar{X}_2$	s_2	$\bar{X}_2 - \bar{X}_1$	t	p	$\bar{X}_1$	s_1	$\bar{X}_2$	s_2	$\bar{X}_2 - \bar{X}_1$	t	p
2	0.392	0.040	**0.547**	0.058	0.156	-3.14	0.0350*	0.448	0.016	**0.490**	0.004	0.041	-3.47	0.0255*
3	0.454	0.032	**0.536**	0.034	0.082	-2.49	0.0671	0.434	0.016	**0.517**	0.024	0.084	-4.04	0.0156*
4	0.433	0.027	**0.556**	0.012	0.123	-5.88	0.0042**	0.482	0.020	**0.522**	0.014	0.040	-2.37	0.0772
5	0.454	0.028	**0.575**	0.012	0.121	-5.67	0.0048**	0.484	0.016	**0.520**	0.016	0.035	-2.19	0.0938
6	0.479	0.044	**0.572**	0.020	0.093	-2.71	0.0534	0.488	0.017	**0.543**	0.010	0.055	-3.92	0.0172*
7	0.503	0.009	**0.560**	0.001	0.057	-8.98	0.0009***	0.507	0.029	**0.529**	0.002	0.022	-1.07	0.3433
8	0.509	0.024	**0.567**	0.017	0.058	-2.83	0.0474*	0.496	0.010	**0.518**	0.019	0.022	-1.40	0.2336
9	0.492	0.016	**0.576**	0.013	0.084	-5.86	0.0042**	0.527	0.015	**0.531**	0.007	0.004	-0.36	0.7350
10	0.475	0.008	**0.566**	0.017	0.091	-6.90	0.0023**	0.536	0.007	**0.538**	0.013	0.002	-0.19	0.8593
11	0.473	0.015	**0.578**	0.008	0.105	-8.87	0.0009***	**0.539**	0.010	0.536	0.011	-0.002	0.24	0.8238
12	0.491	0.010	**0.572**	0.010	0.081	-7.99	0.0013**	**0.550**	0.013	0.545	0.006	-0.005	0.53	0.6255
13	0.484	0.010	**0.591**	0.009	0.107	-11.08	0.0004***	**0.538**	0.007	0.533	0.003	-0.004	0.84	0.4469
14	0.515	0.014	**0.568**	0.006	0.052	-5.03	0.0074**	0.536	0.014	**0.548**	0.003	0.012	-1.15	0.3129
15	0.475	0.022	**0.583**	0.008	0.107	-6.40	0.0031**	**0.558**	0.017	0.555	0.008	-0.003	0.24	0.8195
16	0.485	0.021	**0.585**	0.006	0.100	-6.59	0.0028**	0.542	0.007	**0.561**	0.010	0.019	-2.22	0.0902
17	0.489	0.015	**0.590**	0.022	0.101	-5.40	0.0057**	**0.562**	0.022	0.557	0.009	-0.005	0.27	0.7997
18	0.506	0.035	**0.592**	0.015	0.086	-3.24	0.0315*	**0.558**	0.015	0.550	0.005	-0.008	0.66	0.5441
19	0.493	0.009	**0.589**	0.011	0.096	-9.92	0.0006***	0.543	0.017	**0.553**	0.011	0.010	-0.73	0.5081
20	0.493	0.007	**0.584**	0.009	0.091	-11.54	0.0003***	0.550	0.019	**0.561**	0.006	0.011	-0.82	0.4574
21	0.504	0.020	**0.579**	0.004	0.076	-5.37	0.0058**	**0.569**	0.014	0.560	0.014	-0.009	0.67	0.5398
22	0.497	0.012	**0.576**	0.009	0.079	-7.51	0.0017**	0.559	0.016	**0.564**	0.006	0.005	-0.41	0.7012
23	0.486	0.009	**0.572**	0.022	0.086	-5.09	0.0070**	**0.562**	0.006	0.562	0.012	-0.000	0.04	0.9733
24	0.489	0.001	**0.584**	0.015	0.095	-9.14	0.0008***	0.552	0.008	**0.564**	0.006	0.012	-1.63	0.1794
25	0.471	0.006	**0.567**	0.011	0.096	-10.95	0.0004***	0.548	0.006	**0.564**	0.011	0.016	-1.84	0.1392
26	0.490	0.016	**0.589**	0.019	0.099	-5.72	0.0046**	0.554	0.018	**0.564**	0.011	0.010	-0.67	0.5403
27	0.482	0.013	**0.573**	0.009	0.091	-8.15	0.0012**	0.553	0.010	**0.561**	0.010	0.008	-0.79	0.4720
28	0.488	0.009	**0.585**	0.007	0.097	-12.22	0.0003***	0.552	0.004	**0.567**	0.014	0.015	-1.43	0.2267
29	0.500	0.017	**0.590**	0.002	0.090	-7.50	0.0017**	0.543	0.015	**0.560**	0.003	0.018	-1.68	0.1682
30	0.475	0.010	**0.583**	0.002	0.108	-14.37	0.0001***	**0.558**	0.007	0.557	0.012	-0.001	0.14	0.8980
31	0.478	0.010	**0.584**	0.009	0.105	-11.41	0.0003***	0.557	0.014	**0.568**	0.006	0.011	-1.03	0.3628
32	0.488	0.007	**0.588**	0.006	0.100	-15.40	0.0001***	0.553	0.002	**0.557**	0.003	0.004	-1.61	0.1825
33	0.484	0.013	**0.581**	0.000	0.097	-10.57	0.0005***	0.541	0.009	**0.564**	0.004	0.023	-3.26	0.0311*
34	0.488	0.002	**0.594**	0.010	0.107	-14.57	0.0001***	0.554	0.010	**0.565**	0.013	0.011	-0.97	0.3885
35	0.502	0.011	**0.584**	0.013	0.082	-6.78	0.0025**	0.550	0.002	**0.568**	0.014	0.018	-1.77	0.1521
36	0.481	0.002	**0.578**	0.002	0.097	-59.63	0.0000***	0.550	0.016	**0.573**	0.010	0.023	-1.69	0.1667
37	0.491	0.015	**0.591**	0.009	0.100	-8.17	0.0012**	0.545	0.009	**0.576**	0.005	0.031	-4.18	0.0139*
38	0.476	0.008	**0.580**	0.010	0.105	-12.07	0.0003***	0.550	0.008	**0.565**	0.003	0.014	-2.26	0.0867
39	0.483	0.024	**0.576**	0.008	0.094	-5.26	0.0063**	0.546	0.019	**0.577**	0.005	0.032	-2.32	0.0814
40	0.494	0.007	**0.586**	0.008	0.092	-12.55	0.0002***	0.569	0.016	**0.574**	0.009	0.004	-0.33	0.7560

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

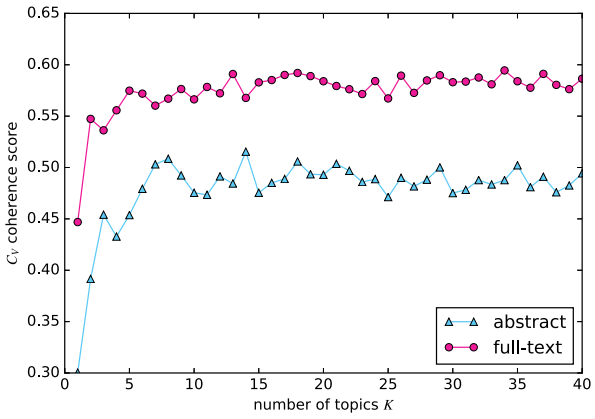
elbow method) at $k = 14$, and the full-text data achieved this at $k = 13$.

B. DS_2 Dataset

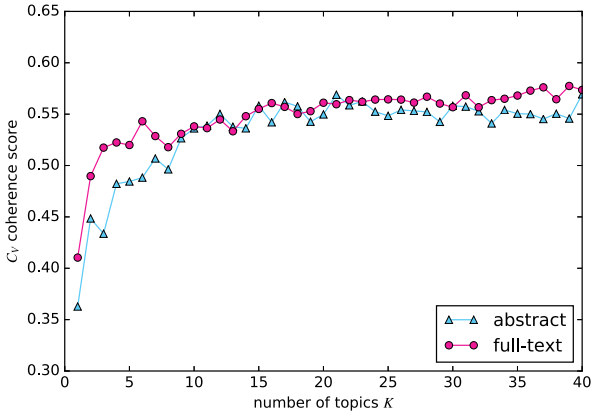
The DS_2 dataset with 15,004 research articles from 12 top-tier fisheries journals show that only 5 LDA models are significantly different at the 5% significance threshold; k = 2, 3, 6, 33, and 37. Looking at the actual coherence scores, most LDA models show a slightly higher coherence score (shown in bold) for full-text data compared to abstract data. However, the large difference in coherence scores and significance levels are not similar to the DS_1 dataset. The LDA model with the optimal coherence score for abstract data is at $k = 17$, and for full-text data at $k = 16$.

C. Human Topic Ranking

Table III shows the results of the human topic ranking by a fisheries domain expert. For an equal comparison, the LDA models with optimal coherence scores were ranked and compared. The LDA model from DS_1 abstract data ($k = 14$)



(a) DS_1 dataset



(b) DS_2 dataset

Fig. 3. Calculated C_V topic coherence score for LDA models with $K = \{1, ..., 40\}$ for (a) DS_1 and (b) DS_2. The coherence score is the mean score for all 3 runs. Scores for DS_1 with 4,417 documents shows that full-text data achieves a higher topic coherence score for all k-values. In contrast, DS_2 with 15,004 documents show similar coherence scores. Individual lines for each run are not shown for clarity.

contains 50% high-quality topics, 36% medium-quality topics, and 14% low-quality topics. In contrast, the LDA model from full-text data ($k = 13$) contains 92% high-quality, 8% medium-quality, and no low-quality topics. DS_2 abstract and full-text data show similar ranking scores; almost 90% high-quality topics with just two topics ranked as medium-quality. Table IV provides an example of high- medium- and low-quality topics, the top 15 words, and the incorrect terms that caused the topics to be ranked lower for DS_1. A two-dimensional inter-topic distance map for the LDA models is displayed in Fig. 4. This two-dimensional representation is obtained by computing the distance between topics [41] and applying multidimensional scaling [42]. It displays the similarity between topics with respect to their probability distribution over words. Furthermore, it shows the topic label that best captures the semantics of the top 15 words. The color

TABLE III
MANUAL TOPIC RANKING FOR DS_1 AND DS_2 DATASETS FOR ABSTRACT AND FULL-TEXT. H = HIGH-QUALITY, M = MEDIUM-QUALITY, AND L = LOW-QUALITY TOPICS.

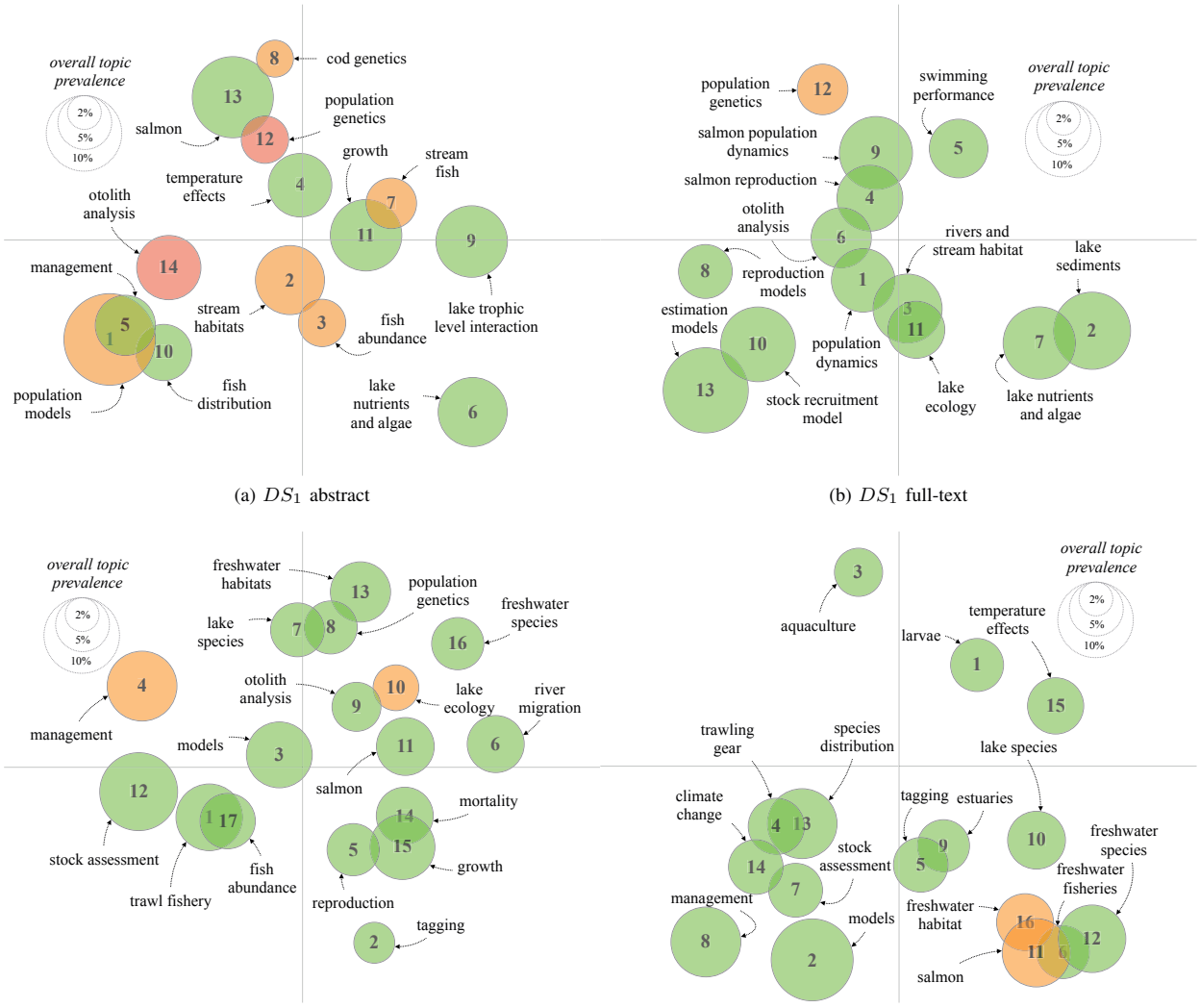
	DS_1		DS_2	
	Abstract	Full-text	Abstract	Full-text
H	7/14 (50.0%)	12/13 (92.3%)	15/17 (88.2%)	14/16 (87.5%)
M	5/14 (35.7%)	1/13 (7.7%)	2/17 (11.8%)	2/16 (12.5%)
L	2/14 (14.3%)	0/13 (0%)	0/17 (0%)	0/16 (0%)

coding indicates the quality of the topics based on human interpretation (see Section III-D for ranking method). It shows that, overall, more high-quality topics are obtained from full-text data than from the abstract counterpart for DS_1, and similar topic rankings are achieved for DS_2.

V. DISCUSSION

The coherence of a topic is based on the topic's top 15 words and shows how strongly pairs of these top 15 words support each other within the corpus. Such an approach, drawing on the philosophical premise that a set of statements or facts is said to be coherent if its statements or facts support each other, informs us about the understandability and interpretability of topics from a human perspective. The LDA models obtained from DS_1 full-text data, compared to DS_1 abstract data, show a higher coherence overall, with the test statistics showing that these differences are significant for all but two LDA models. On the other hand, such significant differences are not present within the DS_2 dataset when comparing abstract and full-text data, although full-text data achieved more topics with a higher coherence score.

Additionally, topic ranking by a fisheries domain expert shows similar, or even greater, improvements in results for the DS_1 full-text data; topics uncovered from full-text data contain 92% high-quality topics compared to 50% high-quality topics from abstract data. The quality of topics from a human perspective was lowered by the inclusion of incorrect terms in the top 15 words. Such terms, however, are not related to the biological, ecological, or socio-ecological meanings of those topics but can be seen as noise terms: *using*, *used*, *use*, *within*, *total*, *two*, *among*, and *within*. There is little to no specific semantic meaning behind these terms, and although they are important in written text, they are less important when uncovering latent semantic structures (i.e. topics) from documents. This issue may be potentially rectified by a part-of-speech (POS) tagger to eliminate the verbs or prepositions that crop up as noise among the topic's top words. However, one should proceed carefully in cases where verbs are important cues for understanding the semantics of the top words. For example, Table IV shows that *fishing* and *feeding* are among the top 15 words, and in these cases, the verbs are important terms that are necessary for the understanding of the semantic context. In such cases, one might proceed with a domain-specific stop word list to prevent such terms from becoming part of the topic-word distribution. A lower ranked topic



(a) DS_1 abstract

(b) DS_1 full-text

(c) DS_2 abstract

(d) DS_2 full-text

Fig. 4. Inter-topic distance map showing a two-dimensional representation (via multi-dimensional scaling) of the latent topics. The distance between the nodes represents the topic similarity with respect to the distributions of words. The surface of the nodes represents the prevalence of the topic within the corpus. Color coding is used to display the topic ranking: green = high-quality topic, orange = medium-quality topic, and red = low-quality topic.

TABLE IV
A SELECTION OF TOPICS FROM DS_1 WITH THE 15 MOST PROBABLE WORDS, TOPIC LABEL, AND RANKING DATA. TEXT IN BOLD INDICATES INCORRECT TERMS.

Dataset	Label	Top 15 words	Ranking
Abstract	fish distribution	fishing, distribution, data, species, areas, catch, abundance, spatial, habitat, model, fishery, effort, fish, water, sea	High
	Population models	model, data, mortality, stock, fish, population, fishing, models, recruitment, cod, estimates, **using**, size, rates, **used**	Medium
	Population genetics	genetic, populations, population, **among**, lake, fish, loci, microsatellite, **two**, structure, diversity, **within**, samples, species, river	Low
Full-text	Salmon population dynamics	salmon, trout, prey, growth, atlantic, temperature, water, rate, juvenile, salmo, feeding, wild, density, food, populations	High
	Population genetics	genetic, populations, population, river, samples, loci, salmon, **among**, dna, atlantic, sample, sea, microsatellite, structure, alleles	Medium

caused by noise terms is not as apparent for full-text data, nor does it seem to hold for abstract data from the DS_2 dataset. Such noise terms seem less of an issue when document frequency, word length, or vocabulary size increases.

Also worth noting is an increased level of detailed topics within DS_1 full-text data (Fig. 4b) compared to DS_1 abstract data (Fig. 4a). For example, the topics *salmon population dynamics* and *salmon reproduction* were uncovered from full-text data, whereas the single topic *salmon* was uncovered from abstract data. Similarly, the topics dealing with lakes are split into three topics (*lake nutrients and algae*, *lake sediments*, and *lake ecology*) from full-text data, compared to two (*lake tropic level interaction* and *lake nutrients and algae*) from the abstract data. Lastly, the topics dealing with models were split into three (*estimation models*, *stock assessment models*, and *reproduction models*) from full-text data in contrast to an overarching *population model* topic from abstract data. Such a clear difference between low and high granularity topics is not present within the DS_2 dataset. Although the differences in word length and vocabulary size exists, similarly to DS_1, it seems that a higher number of documents makes up for these differences in granularity. A comparison between other LDA models (not presented) shows similar granularity between abstract and full-text for DS_2. Although the article's abstract aims to provide a complete but succinct description of the whole paper, it is often restricted by a limitation on the number of words. Such word limitation, with a relatively small number of documents, has practical effects on the level of detail (i.e. granularity) of uncovered LDA topics.

Besides topic coherence, topic ranking, and the level of detail, Fig. 4 shows a number of uncovered topics that are present in abstract data but absent in full-text data. Within DS_1 for example, the topics *temperature effects*, *cod genetics*, *management*, and *fish abundance* were not found within full-text data, and neither were related topics showing semantic resemblance to these absent topics. Although we identified similar and detailed topics, there remains an inconsistency between some uncovered topics from both datasets. Knowing that abstracts were retrieved from full-text articles and are thus, in essence, a subset of the full-text data, one might question why these differences exist. One reason might be that manual topic labeling is limited to the subjectivity inherent in human interpretation and an analysis of the topics by others could yield opposite results, explaining away any differences between the two datasets. On the other hand, topic labeling is usually performed by inspection of the topic words with the highest probabilities (top 10 or 15). Such an approach might up-weight terms that have high probability under all topics. Other approaches to identify the terms that best describe a topic exist (e.g. [7], [43]) and could yield different results. Finally, abstract data, being restricted by the limited number of words, fail to adequately convey the heterogeneity of research ideas or topics that are part of a document. Uncovered latent topics might thus not completely resemble the document collection and, as a result, provide a limited or even incorrect view of the underlying thematic structure.

VI. Conclusion

In this paper, we presented a comparison between topic coherence scores and human topic ranking when creating LDA topics from abstract and full-text data. Two datasets were compared, DS_1 consisting of a single fisheries journal with 4,417 scientific research articles that span 20 years of scientific output, and DS_2 consisting of 12 fisheries journals, 15,004 articles, and span 16 years of research. The two types of data, abstract and full-text, combined with two different datasets, a single journal and a set of journals, allow for comparison on a variety of characteristics, such as document length, document frequency, and vocabulary size. Topics were statistically compared by adopting the C_V coherence measure that shows the highest correlation with all available human topic-ranking data. Furthermore, the LDA models with the optimal coherence scores were manually inspected and ranked by a fisheries domain expert.

Our results show that uncovering LDA models from a single journal with, relatively speaking, a low number of documents are very prone to noise terms that crop up into the topic's top words—the words that are often used to capture the semantics of the topic—for abstract data. Such noise terms require special attention when dealing with abstract data with, e.g. an increased cleaning phase, POS filtering, or a domain-specific stop word list. Our results show that full-text data seem less affected by such words, thus increasing the coherence and manual topic ranking. On the other hand, increasing the number of document (e.g. DS_2) results in fewer noise terms, thus an improvement in coherence and human topic ranking for both abstract and full-text data. Furthermore, on a small dataset (e.g. DS_1) abstract topic distributions capture more broad topics, with full-text topics achieving more fine-grained results. These differences in detail are not present for bigger datasets containing a higher number of documents, regardless the choice for abstract or full-text data.

We identified a number of topics that were uncovered from abstract data but were absent among the topics uncovered from full-text data. A detailed analysis of the reasons behind these differences would yield interesting results and would be a possible direction for future research.

Acknowledgment

The authors are grateful to Charlotte Teresa Weber for ranking and labeling the topics, and to Melania Borit, Lia ní Aodha, and Bruce Edmonds for improving earlier versions of this article. This research was funded by the project SAF21, Social Science Aspects of Fisheries for the 21^{st} Century. SAF21 is a project financed under the EU Horizon 2020 Marie Skłodowska-Curie (MSC) ITN ETN program (project 642080).

References

[1] P. O. Larsen and M. von Ins, "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index," *Scientometrics*, vol. 84, no. 3, pp. 575–603, sep 2010.

[2] K. W. Boyack and R. Klavans, "Creation of a highly detailed, dynamic, global model and map of science," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 670–685, apr 2014.

[3] A. Srivastava and M. Sahami, *Text mining: Classification, clustering, and applications*. CRC Press, 2009.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, sep 1990.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. New York, New York, USA: ACM Press, 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[7] D. M. Blei and J. D. Lafferty, "Topic Models," *Text Mining: Classification, Clustering, and Applications*, pp. 71–89, 2009.

[8] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Supplement 1, pp. 5228–5235, apr 2004.

[9] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 03, pp. 267–297, jan 2013.

[10] T. Rusch, P. Hofmarcher, R. Hatzinger, and K. Hornik, "Model trees with topic model preprocessing: An approach for data journalism illustrated with the WikiLeaks Afghanistan war logs," *The Annals of Applied Statistics*, vol. 7, no. 2, pp. 613–639, jun 2013.

[11] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl, no. suppl 1, pp. 5220–5227, apr 2004.

[12] C. J. Gatti, J. D. Brooks, and S. G. Nurre, "A Historical Analysis of the Field of OR/MS using Topic Models," *arXiv.org*, vol. stat.ML, oct 2015.

[13] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transportation Research Part C: Emerging Technologies*, vol. 77, no. June, pp. 49–66, apr 2017.

[14] M. J. Westgate, P. S. Barton, J. C. Pierson, and D. B. Lindenmayer, "Text analysis tools for identification of emerging topics and research gaps in conservation science," *Conservation Biology*, vol. 29, no. 6, pp. 1606–1614, 2015.

[15] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 363–371.

[16] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 424–433.

[17] J. M. Alston and P. G. Pardey, "Six decades of agricultural and resource economics in Australia: an analysis of trends in topics, authorship and collaboration," *Australian Journal of Agricultural and Resource Economics*, vol. 60, no. 4, pp. 554–568, oct 2016.

[18] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Association for Computational Linguistics, 2013, pp. 13–22.

[19] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. New York, New York, USA: ACM Press, 2015, pp. 399–408.

[20] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring Topic Coherence over Many Models and Many Topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, no. July. Association for Computational Linguistics, 2012, pp. 952–961.

[21] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[22] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, aug 1954.

[23] W. L. Buntine, "Operations for Learning with Graphical Models," *Journal of Artificial Intelligence Research*, vol. 2, pp. 159–225, nov 1994.

[24] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, vol. 55, no. 4, apr 2012, pp. 77–84.

[25] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed inference for latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, vol. 20, 2007, pp. 1081–1088.

[26] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, p. 569.

[27] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, mar 2006.

[28] Y. W. Teh, D. Newman, M. Welling, and D. Neaman, "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, 2007, pp. 1353–1360.

[29] C. Wang, J. Paisley, and D. M. Blei, "Online Variational Inference for the Hierarchical Dirichlet Process," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011, pp. 752–760.

[30] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On Smoothing and Inference for Topic Models," *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, no. Ml, pp. 27–34, may 2012.

[31] D. Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan and Mimno, "Evaluation Methods for Topic Models," in *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1105–1112.

[32] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 288–296.

[33] D. Newman, J. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, no. June. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 100–108.

[34] I. Douven and W. Meijs, "Measuring coherence," *Synthese*, vol. 156, no. 3, pp. 405–425, 2007.

[35] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," in *Proceedings of German Society for Computational Linguistics (GSCL 2009)*, 2009, pp. 31–40.

[36] I. Jarić, G. Cvijanović, J. Knežević-Jarić, and M. Lenhardt, "Trends in Fisheries Science from 2000 to 2009: A Bibliometric Study," *Reviews in Fisheries Science*, vol. 20, no. 2, pp. 70–79, 2012.

[37] S. Syed, M. Spruit, and M. Borit, "Bootstrapping a Semantic Lexicon on Verb Similarities," in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, vol. 1, no. Ic3k. SCITEPRESS - Science and Technology Publications, 2016, pp. 189–196.

[38] N. Evangelopoulos, X. Zhang, and V. R. Prybutok, "Latent Semantic Analysis: five methodological recommendations," *European Journal of Information Systems*, vol. 21, no. 1, pp. 70–86, jan 2012.

[39] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[40] M. D. Hoffman, D. M. Blei, and F. Bach, "Online Learning for Latent Dirichlet Allocation," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 2010, pp. 856–864.

[41] J. Chuang, D. Ramage, C. D. Manning, and J. Heer, "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis," in *ACM Human Factors in Computing Systems (CHI)*, 2005, pp. 443–452.

[42] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70.

[43] Z. Tang and J. Maclennan, *Data Mining With SQL Server 2005*. Wiley, 2005.