# Topic Analysis of Climate-Change News

Sudarshan S. Chawathe
*School of Computing and Information Science & Climate Change Institute*
*University of Maine*
Orono, Maine 04469-5711, USA
chaw@eip10.org

***Abstract*—This paper explores the application of computational methods to the analysis of the large and growing corpus of news articles and related data on climate change. Topics are analyzed using Latent Dirichlet Allocation and methods customized to specific news sources that take advantage of keywords and other metadata that may be present. Results of this method on news articles drawn over several months are presented.**

***Index Terms*—Climate Change, News, Topic Modeling, Machine Learning.**

## I. Introduction

Understanding public perceptions of diverse climate-change issues is important to policymakers and others devising adaptation strategies. News articles in the international, national, and local press provide valuable summaries of such perceptions and issues. However, the volume of these articles is too large to allow direct human examination of anything more than a small fraction by any individual or research group. This situation motivates methods for computational analysis of news and related data sources.

There are diverse techniques that may be applied to corpora of news or other textual documents, and a large body of work spanning several decades. The focus of this paper is on a relatively recent technique for topic modeling using Latent Dirichlet Allocation (LDA) [1]. The goal is to develop a framework and implementation for analyzing topics in news articles that permits rapid comprehension of a large corpus of documents, far beyond what may be achieved by solely human effort. The *main question addressed by this paper* is: How well can topic modeling in general, and LDA in particular, be applied to the problem of analyzing news articles in general, and news on climate change in particular?

The main *contributions* of this paper may be summarized as follows:

- It motivates and develops a framework for topic analysis of climate-change news.
- It outlines the design and implementation of a prototype system for such analysis, and reports on early experience with it on real data interacting with archival services.
- It summarizes results from an experimental evaluation of this system on news articles related to climate change drawn from several months of the New York Times archives.
- Based on the experimental results and early experience with the prototype, it outlines some promising avenues for future work.

The main *results* may be summarized as follows:

- A combination of diverse quantitative evaluation metrics and selective qualitative examination of the underlying data guided by these metrics provides an effective toolbox for judging the quality of topic models and for using the topics to understand the corpus (e.g., Fig. 14 and Table II, Pages 6 and 7).
- Streamlining data acquisition and preparation are key steps to enabling automated topic modeling and semi-automated comprehension of a large corpus of news articles.
- Carefully crafted visual representations of topics and documents are important for going beyond the numerical metrics.
- Metadata accompanying news articles is a noisy but promising source for refining topics and, in particular, for assigning human-meaningful descriptions to them.

*Paper outline:* Section II formalizes the problem that was informally described above and provides details on both the document model and the generative process conceptually assumed by LDA. The prototype system implementation is described in Section III. The important issue of evaluating the quality of topic models generated by LDA or other methods (including human) is addressed in Section IV. Experimental results are summarized in Section V. An important aspect of this work is the presentation of both numerical results (e.g., topic model metrics) and illustrative qualitative results and samples from a real and important corpus. Related work is addressed in Section VI. Section VII provides a brief summary and topics of ongoing work.

## II. Latent Dirichlet Allocation

Let $D$ denote the corpus of documents. Each document is modeled as a bag (multiset) of words: $d = \{w_1, w_2, \ldots, w_n\}$. The number of occurrences of a word in a document is a key feature, where each word is drawn from a finite dictionary or vocabulary $W$. We assume that the corpus is associated with an unknown set of topics: $T = \{t_1, t_2, \ldots, t_k\}$. Each topic $t_i$ determines probability distribution on words.

$$P(w) = \sum_{t \in T} P(w \mid t) \cdot P(t)$$

978-1-7281-3783-4/20/$31.00 ©2020 IEEE

The main idea behind LDA is an assumed conceptual generative process for documents based on topics (unknown). A document, viz., a bag of words, is conceptually generated by repeatedly selecting a topic, using some probability distribution over topics, followed by selecting a word for that topic using a probability distribution of words for that topic.

The *main task* is now that of inferring $T$ given the corpus $D$. This task is conceptually easy: All we need to do is to evaluate for all possibilities and use Bayes Theorem. However, this conceptual simplicity is deceptive because the method it implies is completely infeasible in practice due to numbers that are beyond astronomically large. It is therefore a hard problem to solve but there is a long history of work on it and many workable solutions. A popular option, and one that is implemented in the Mallet system [2] used in the implementation and experimental study, is to use Gibbs sampling.

## III. Prototype System Implementation

The prototype topic analysis system is implemented on the Java Virtual Machine (JVM) platform using Kawa Scheme [3]. Experiments are conducted using OpenJDK [4] on a computer running Debian 9 GNU/Linux. The core topic analysis tasks are performed by using the Mallet toolkit [2]. Interactions with Web APIs are conducted using the *OkHttp* Java client accessed from Kawa.

For the study reported here, 15 months of New York Times articles matching the search phrase "climate change" were retrieved from the ProQuest archive using an institutional subscription. Unfortunately, there was no suitable API available and so the implementation must parse the article details from a text file representing the aggregated search results. Articles retrieved from such archives often have some structure and metadata that may be profitably used for topic analysis. Unfortunately, such structure and metadata is often in peculiar formats which presents a challenge for parsing and assimilating. The current implementation uses a simple implementation of an LL(1) parser that was found to be effective at parsing the archive. The result of such parsing is stored as a Lisp S-expression, one per document.

The parsed articles are in turn used to generate plain-text documents that form the input to the topic analysis module. For the experiments reported herein, such a plain-text document is generated from the parsed representation simply by concatenating the title and full text. An interesting avenue of ongoing work is studying the effect of including some of the other fields from the parsed document data. The topics and keywords as identified by such metadata need not be consistent with those identified by methods such as LDA. However, the metadata may be used to assign human-understandable topic descriptions to the topics identified by fully automated methods.

The core of the LDA computations are performed using the Mallet toolkit via its Java API. In particular, the document archive is split into subsets corresponding to calendar months of the articles' publication dates and each subset is analyzed separately. Mallet is also used for computing several metrics of topic quality (Section IV). Results are plotted interactively using a combination of Scheme and gnuplot.

Another module of the prototype system uses the parsed representation of articles and topics to generate graphical summaries such as the one Fig. 2.

## IV. Evaluating Topic Models

Determining the quality of the topic model computed by an automated method (or by a human) is a challenging task because there is often no consensus among experts on how a corpus may best be mapped to topics. Nevertheless, there has been considerable work on discovering objective measures of topic model quality and some combination of these measures provides a reasonable picture of a topic model's quality or, perhaps more importantly, suitability for a given purpose.

Let $C(w, d, t)$ denote the number of occurrences (*counts*) of the word $w$ in document $d$ that are assigned to topic $t$. We use a convention that omitting one or more of the arguments of $C$ denotes a value obtained by summing over all possible values of the omitted arguments. The identities of the omitted arguments are inferred from the parameter names: $w$ for word, $d$ for document, $t$ for topic. Thus, for example,

$$C(d,t) = \sum_{w \in W} C(w,d,t)$$

where $W$ is the set of all words in the corpus. Then the probability of a document $d$ given a topic $t$ is

$$\hat{P}(d \mid t) = \frac{C(d,t)}{C(t)}$$

where we use $\hat{P}$ to denote an estimate. Following standard definitions, the *document entropy* is

$$H(t) = -\sum_{d \in t} P(d \mid t) \log P(d \mid t)$$

A topic with low entropy is more likely to be heavily represented in a few documents whereas one with high entropy is more likely to be distributed more broadly over the corpus.

Instead of focusing on the documents for a topic, we may consider the co-occurrence of words, leading to the *coherence* metric, which quantifies the degree to which the most frequently occurring words in a topic co-occur. In more detail, for words $w$ and $x$, let $D(w, x)$ denote the number of documents in the corpus that contain at least one occurrence of each word word (regardless of topic). Similarly, let $D(w)$ denote the number of documents with at least one occurrence of the word $w$. Further, let $w(t, i)$ denote the $i$-th highest ranked word for $r = 1, 2, \ldots, k$ for some fixed $k$ which represents the number of top-ranked words being considered, with lower values of $r$ denoting higher rank. (In the experiments, $k = 20$.) Then the coherence is

$$Coh_k(t) = \sum_{r=1}^{k} \sum_{s=r+1}^{k} \log \left( \frac{D(w(t,i), w(t,j)) + \beta}{D(w(t,i))} \right)$$

where $\beta$ is a constant added to avoid taking the logarithm of zero in case the main term in the numerator is zero.

A related word-focused metric is the degree to which the most frequent words of are exclusive to that topic, i.e., the degree to which they do not appear as the most frequent words of other topics. This *exclusivity* metric is computed as

$$Exc_k(t) = \frac{1}{k}\sum_{i=1}^{k}\frac{P(w(t,i) \mid t)}{\sum_{t' \in T} P(w(t,i) \mid t')}$$

where, as before, $k$ is the number of ranked words and $T$ is the set of all topics.

Another kind of metric quantifies the degree to which the computed topics are different from a baseline distribution. The difference is quantified using the *relative entropy*, also known as the *Kullback-Leibler divergence*, which is defined as follows for discrete probability distributions $P_1$ and $P_2$:

$$H_{rel}(P_1 \parallel P_2) = -\sum_x P_1(x)\log\left(\frac{P_1(x)}{P_2(x)}\right)$$

where the summation is over all possible values of $x$. Different choices for the baseline distribution ($P_2$ in the definition above) yield different metrics of this kind.

Using the uniform distribution as the baseline, we obtain the *uniform distance* metric for a topic $t$:

$$\begin{aligned}\delta_U(t) &= H_{rel}(P \parallel U)\\ &= -\sum_{w \in W} P(w \mid t)\log\left(\frac{P(w \mid t)}{U(w \mid t)}\right)\\ &= -\sum_{w \in W} P(w \mid t)\log\left(|W| \cdot P(w \mid t)\right)\end{aligned}$$

where $W$ is the set of all words in the corpus and the last equality follows from the uniform distribution $U(w \mid t) = 1/|W|$.

Since the occurrence of words in a corpus is likely to be significantly different from uniform, an alternative to $\delta_U$ uses the actual distribution of words in the corpus as a baseline. The resulting *corpus distance* metric for a topic $t$ computed similarly, replacing $U$ above with the corpus distribution for words.

## V. Experimental Results

Using the prototype implementation described earlier, articles matching the phrase "climate change" were downloaded for a period of 15 months and analyzed for topics, metadata, and keywords.

The remaining results are summarized here using data from only three representative months, for brevity. Fig. 1 depicts a heat map of topics (horizontal axis) by documents (vertical axis). The keywords for each of these topic are listed in Table I.

Figs. 3 to 13 summarize the results of computing several of the metrics for evaluating topic models (Section IV) for a representative month of data (2018-02). Fig. 14 summarizes the effect of changing a key parameter, the number of computed topics (9, 12, 15), on the document-entropy metric for the data of a representative month (2018-02). It is worth noting that in all three cases there is is a clear outlier with high entropy,

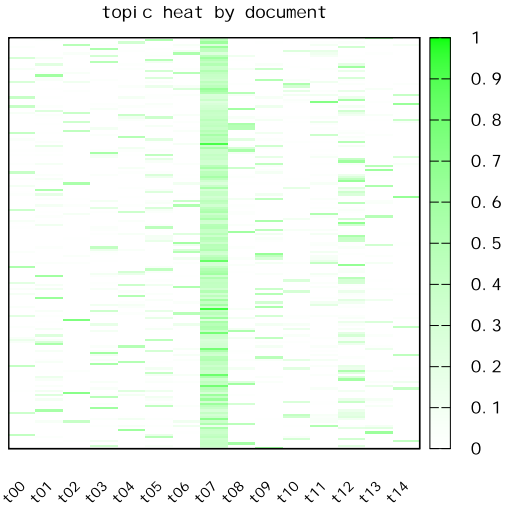


Fig. 1. Heat map of topics by document for 2019-03. See Table I.

identified as topic 7, 5, and 8, in the 9-topic, 12-topic, and 15-topic charts, respectively. The top-ranked keywords for these topics from the experiments are summarized in Table II (upper part). It is immediately clear that this topic is one focused on mostly noise words related to the publication and company and likely found in all documents in this archive, and thus may be omitted from further study. (This observation is consistent with the discussion of entropy in Section IV where high topic entropy is characterized as the topic being distributed broadly over documents in the corpus.) A similar summary for the lowest-entropy topic in each case is also included in Table II (lower part). Although the precise topic semantics are a bit unclear, it is clear that the three topics (for the 9-, 12-, and 15-topic cases) are very similar based on the top keywords. The prototype system implementation facilitates study of this kind, which is useful not only for interpreting the topics but also for iteratively refining the topic-modeling parameters.

## VI. Related Work

Topic modeling has been applied to study the global climate policy debate [5]. That work analyzes 677 articles from two newspaper sources by hand-coded political claims before automated topic modeling. The resulting topics are also validated by human experts using a qualitative reading of the top ten words and documents in each topic. A combination of LDA and discourse analysis has been used to explore debates on social policy related to ethanol production in Brazil [6]. A Markov chain Monte Carlo (MCMC) algorithm for inference has been applied to scientific articles to identify hot topics [7]. Topic modeling has also been applied to organize scientific literature using predefined domain-specific ontologies [8].

An alternative to LDA-based topic modeling is provided by Hierarchical Latent Tree Models (HLTMs) which model document collections using trees [9]. Nodes at the lowest level represent occurrence of words in a document while those at other levels are binary variables modeling co-occurrence of the concepts of the immediately lower level. Such models are

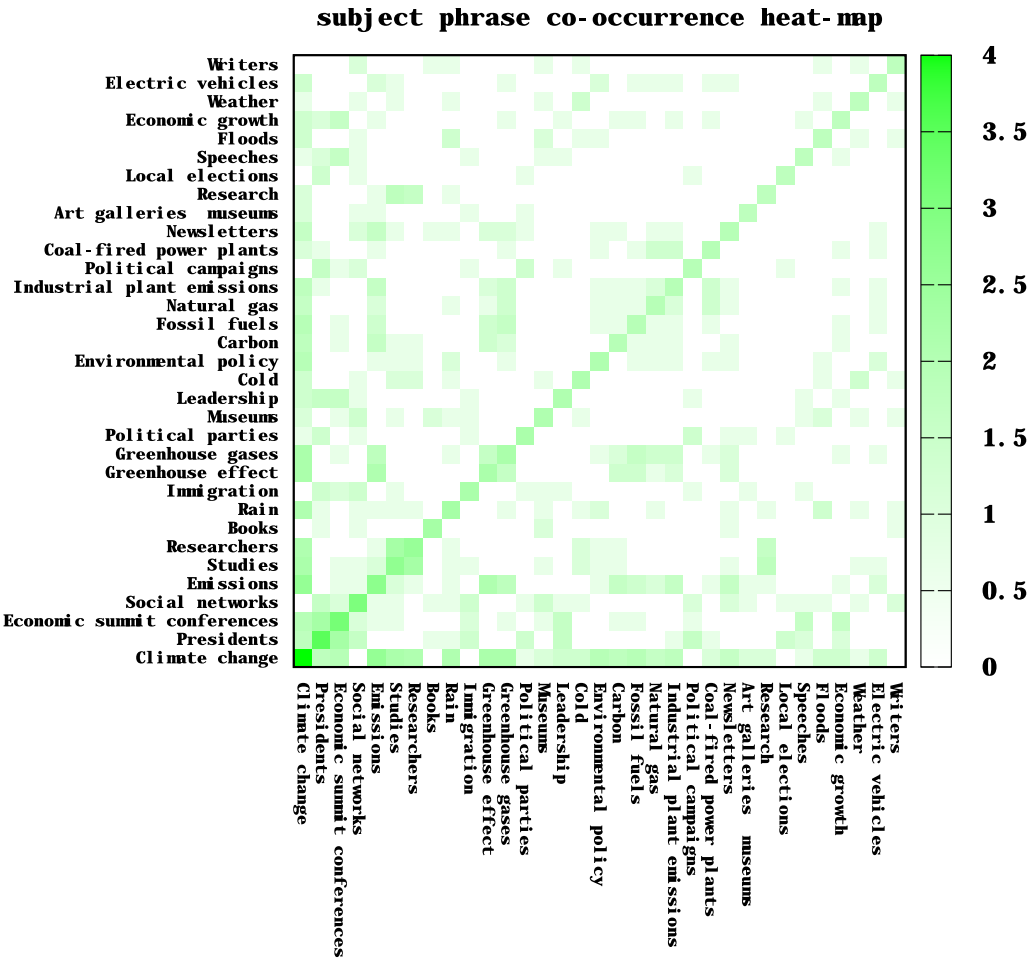


Fig. 2. Subject-phrase co-occurrence heat-map for 2018-01, using a log (base 10) scale for occurrences.

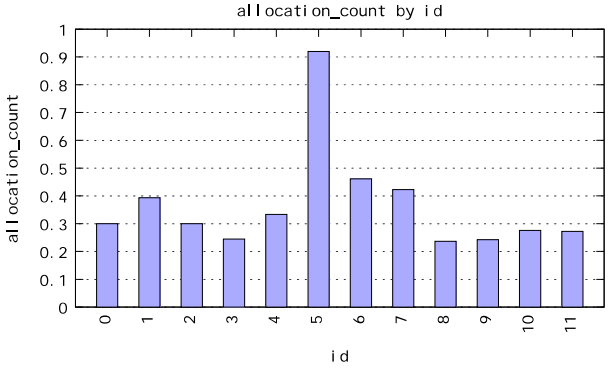


Fig. 3. Topic allocation counts

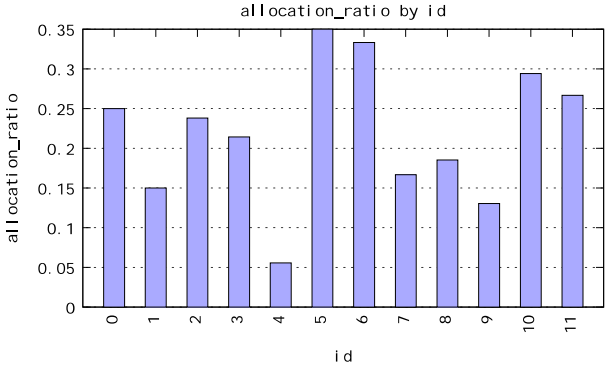


Fig. 4. Topic allocation ratios

potentially more conducive to discovering meaningful topic hierarchies than are LDA models.

Evaluating topic models is a challenging task due to computational and other difficulties. As has been noted in prior work [10], a natural metric is the probability (as computed by the model) of documents from the corpus that are held out for testing. However, exact computation of this probability is intractable, which forces the use of estimation methods. That work outlines the disadvantages of some of these methods and proposes two alternative ones, one of which is implemented in MALLET [2].

In addition to information gleaned from news and other media sources that are the focus of this work, agronomic condition testing systems [11] and big data [12] are valuable

TABLE I
TOPIC KEYWORDS FOR 2019-03. SEE FIG. 1.

| ID | weight | keywords |
|---|---|---|
| 0 | 0.05456 | mueller investigation business trump oil report public fund executives robert companies justice counsel billion saudi muellers iii special safety election |
| 1 | 0.09943 | trump budget congress trumps president house administration federal senate space science special pence billion emergency spending national wall programs naics |
| 2 | 0.02448 | theater march april game season play show netflix series ring previews thrones starts streaming opens drama power p.m plays stars |
| 3 | 0.08341 | climate plastic ice warming weather heat scientists bags cold science species change global waves world paper waste university bag found |
| 4 | 0.05359 | china trump nations united european world chinese international iran france union president europe foreign malpass bank french economic american chinas |
| 5 | 0.11837 | green deal gas energy emissions climate percent carbon greenhouse fossil environmental oil beal power housing plan renewable natural reduce fuels |
| 6 | 0.06024 | book books story fiction science horror read wall future true author history john wallace-wells life stories writes writers literature human |
| 7 | 1.17815 | york times publication company online climate naics states document change people united subject type title proq_ss&genre &issue &title accountid &spage |
| 8 | 0.06879 | water river floods flooding missouri weather flood levees snow rain levee iowa california rivers nebraska corps storm iditarod town south |
| 9 | 0.06261 | briefing zealand brexit minister naics boeing heres u.s prime president friday trump evening european parliament attack israel day vote briefing&author |
| 10 | 0.05012 | orourke beto rural immigration city black milwaukee iowa race convention harris wisconsin paso texas garden manhattan border christian greens housing |
| 11 | 0.0988 | school high women students people life work world legacy kids men class things society make dont children man mental young |
| 12 | 0.18664 | democratic democrats trump president presidential party campaign senate care political candidates politics health voters senator house republicans sanders vote policy |
| 13 | 0.04842 | trump children party review donald asylum magazine pope david op-ed michelle republican politics writer roberts parents ross truth google ive |
| 14 | 0.02572 | museum food wines art chefs australian design puerto australia local cooking wine west island rico restaurants river farmers valley restaurant |

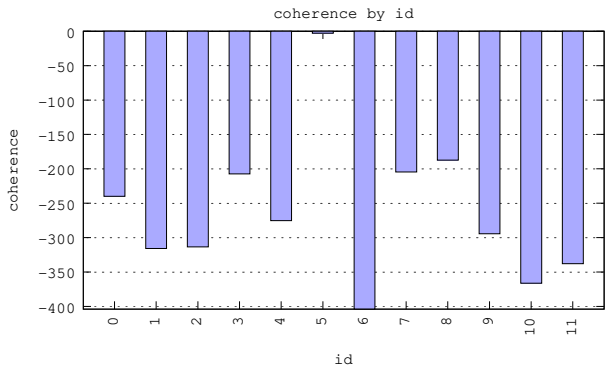


Fig. 5. Topic coherence

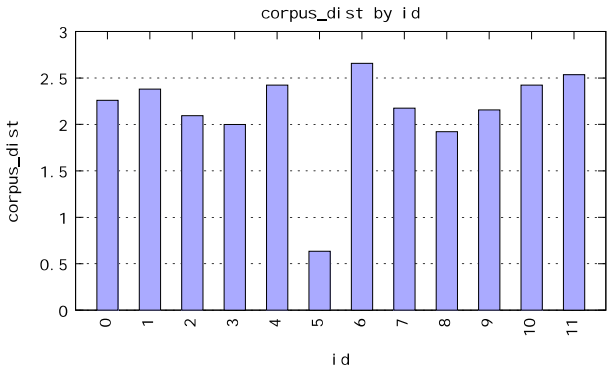


Fig. 6. Topic corpus distances

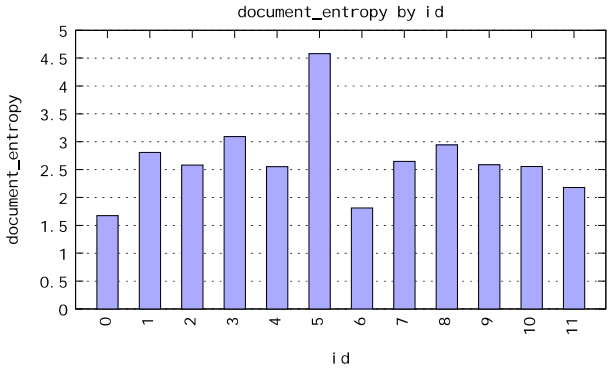


Fig. 7. Topic document entropies

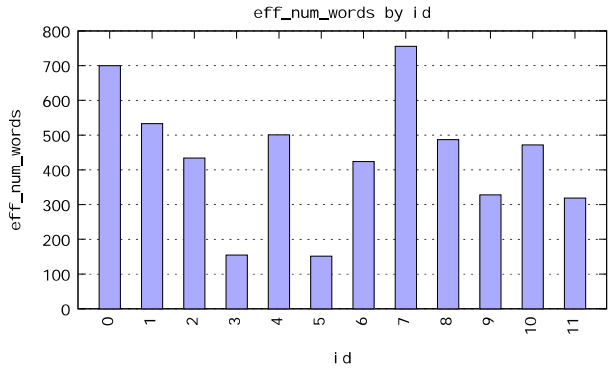


Fig. 8. Topic effective number of words

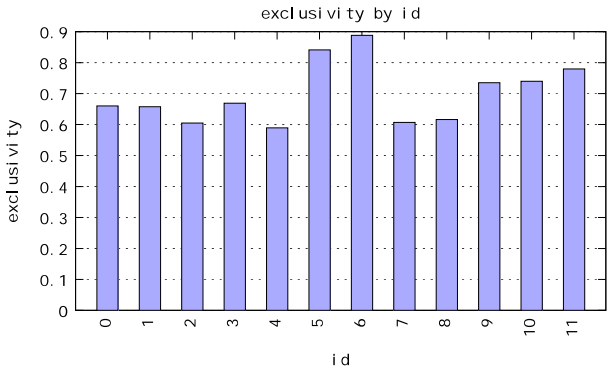


Fig. 9. Topic exclusivity

TABLE II
HIGHEST- AND LOWEST-ENTROPY TOPICS FROM FIG. 14.

| num-topics | ID | Dirichlet param. | top keywords |
|---|---|---|---|
| 9 | 7 | 0.85135 | york times publication company online naics states document united people change climate year feb subject global title amp;title copyright amp;id |
| 12 | 5 | 0.89114 | york times publication company online naics states climate document united change people global feb subject year title amp;issue amp;title copyright |
| 15 | 8 | 0.88996 | york times publication company online naics states climate document change united people global feb subject title amp;title year amp;issue https://search.proquest.com/docview |
| 9 | 8 | 0.04949 | foster climate im change ward california valve children criminal angeles ski kids actions child action time oil skiing turners defense |
| 12 | 0 | 0.03399 | foster art ward valve museum center mercer change movement family criminal years minnie im turners street climate work defense manhattan |
| 15 | 13 | 0.04298 | foster climate change ward california valve children im angeles criminal actions family movement child washington los turners felt action judge |

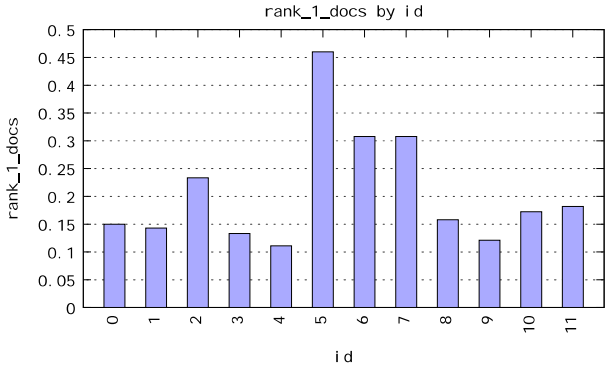


Fig. 10. Topic rank-1 documents

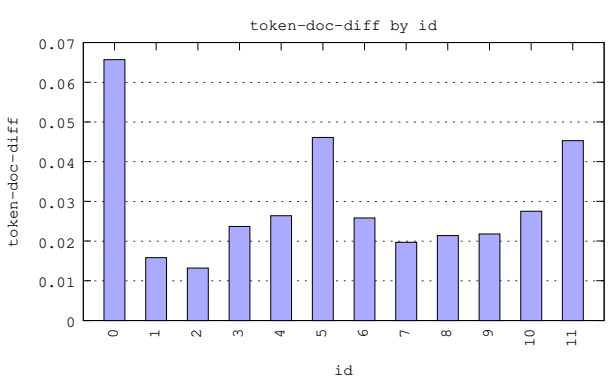


Fig. 11. Topic token-document differences

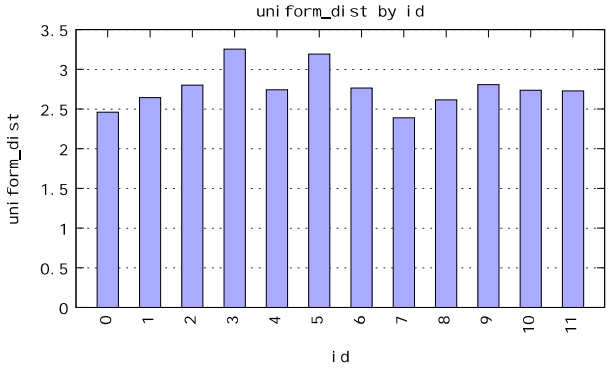


Fig. 12. Topic uniform distances

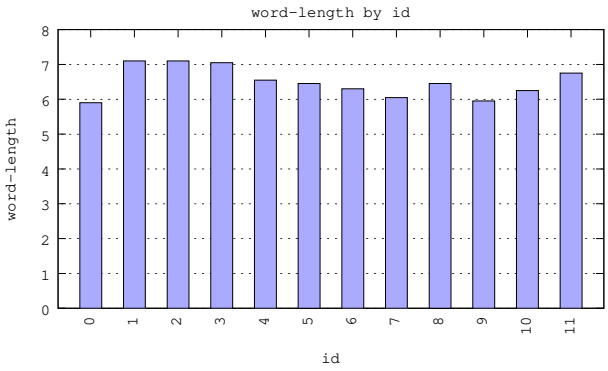


Fig. 13. Topic word lengths

for governments and other decision makers in the context of climate change. High-throughput clustering methods [13], [14] may be applied to high-volume streams of textual news from traditional and social media. Detecting fake information is an especially pressing need for social media sources [15] and recent methods based on deep learning [16] may be used. When user authentication is used to control data quality, methods that avoid reliance on passwords [17] enhance security. Moving from text to audio signals, methods such as those for analyzing emotional prosody [18] provide attributes for further analysis. In a distributed deployment, recent work on hierarchical and distributed machine learning [19] may be used to achieve performance and energy use goals.

## VII. CONCLUSION

Automated or semi-automated analysis of topics in the climate-change discourse as found in the news media is an important source of information that may guide policies for adaptation to climate change in a manner that is sensitive to the needs and opinions of local populations. However, such analysis is challenging due to not only the usual difficulties associated with topic analysis (such as efficient implementation, determining the number of topics, and assigning meaningful

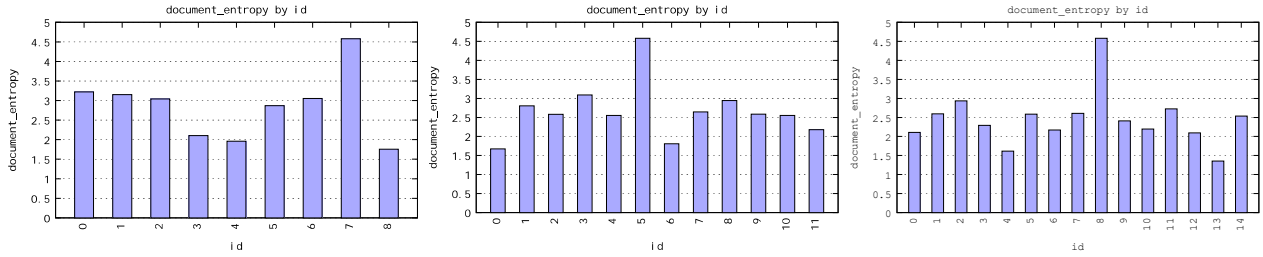


Fig. 14. Topic entropy with varying number of computed topics: 9 (left), 12 (middle), and 15 (right), for New York Times articles matching "climate change" in February 2018. See also Table II.

semantics to topics identified by bags of words) but also due to system implementation aspects related to restricted and nonstandard APIs by which much data must be accessed.

This paper reported work on a prototype implementation of such a system along with some representative results based on extracting news articles related to climate-change over a 15-month period. The data and metadata for these articles was analyzed using well established methods such as LDA (Latent Dirichlet Allocation) as well as methods customized to the specific corpus. An interesting future aspect of this work is the use of such metadata to seed the process of assigning human-meaningful descriptions to automatically identified topics.

Continuing work is expanding the scope of this work by incorporating not only larger volumes of data but also greater variety by using diverse sources. The disparate metadata provided by such sources is both a challenge (for integration) and an opportunity (for seeding topics), and is a focus of ongoing work. Further, the prototype system is being extended to make it easier extract increasingly detailed document information from high-level summaries, to reduce the time and effort required to make observations similar to those made for the sample results near the end of Section V.

## ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation grants 1027960, 1142007, and 1848747. The presentation benefited from detailed comments in the reviews.

## REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.

[2] Andrew Kachites McCallum, "MALLET: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[3] Per Bothner, "The Kawa Scheme language," Manual for version 3.0. https://www.gnu.org/software/kawa/, 2017.

[4] Oracle Corp. *et al.*, "OpenJDK," http://openjdk.java.net/, 2017.Accessed on 16-07-2019.

[5] Veikko Oskdri Eranti, Anna Kristiina Kukkonen, and Tuukka Salu Santeri Ylä-Anttila, "Topic modeling the global climate policy debate," in *Proceedings of the International Conference on Computational Social Science ($IC^2S^2$)*, Helsinki, Finland, Jun.8-11 2015, poster.

[6] L. L. Benites-Lazaro, L. Giatti, and A. Giarolla, "Topic modeling method for analyzing social actor discourses on climate change, energy and food security," *Energy Research & Social Science*, vol. 45, pp. 318–330, 2018, special Issue on the Problems of Methods in Climate and Energy Research. http://www.sciencedirect.com/science/article/pii/S2214629618307990

[7] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proc Natl Acad Sci U S A*, vol. 101 Suppl 1, no. Suppl 1, pp. 5228–5235, Apr. 2004, 14872004[pmid]. https://www.ncbi.nlm.nih.gov/pubmed/14872004

[8] Jennifer Sleeman, Tim Finin, and Milton Halem, "Ontology-grounded topic modeling for climate science research," in *Emerging Topics in Semantic Technologies. ISWC 2018 Satellite Events*. AKA Verlag, Berlin, Oct. 2018.

[9] Peixian Chen, Nevin L. Zhang, Tengfei Liu, Leonard K. M. Poon, and Zhourong Chen, "Latent tree models for hierarchical topic detection," *CoRR*, vol. abs/1605.06650, 2016. http://arxiv.org/abs/1605.06650

[10] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1105–1112.

[11] W. Krause, A. Jarrett, J. Bertish, and C. Jaiswal, "Field agronomic condition test (F.A.C.T.) environmental sensing: The future of agricultural and conservation IOT," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, Nov. 2018, pp. 196–202.

[12] M. Pannu, B. Gill, W. Tebb, and Kai Yang, "The impact of big data on government processes," in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Oct. 2016, pp. 1–5.

[13] K. Rabbi, Q. Mamun, and M. R. Islam, "Dynamic feature selection (DFS) based data clustering technique on sensory data streaming in ehealth record system," in *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, Jun. 2015, pp. 661–665.

[14] H. Bhaumik, B. Chakraborty, A. Mukherjee, S. Bhattacharyya, and M. Chattopadhyay, "Towards reliable clustering of english text documents using correlation coefficient," in *2014 International Conference on Computational Intelligence and Communication Networks*, Nov. 2014, pp. 530–535.

[15] A. Campan, A. Cuzzocrea, and T. M. Truta, "Fighting fake news spread in online social networks: Actual trends and future research directions," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 4453–4457.

[16] A. Verma, V. Mittal, and S. Dawn, "FIND: Fake information and news detections using deep learning," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Aug. 2019, pp. 1–7.

[17] Garima Bajwa, Ram Dantu, and Ryan Aldridge, "Pass-pic: A mobile user authentication," in *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, May 2015, p. 195.

[18] N. Ang, D. Bein, D. Dao, L. Sanchez, J. Tran, and N. Vurdien, "Emotional prosody analysis on human voices," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2018, pp. 737–741.

[19] A. Thomas, Y. Guo, Y. Kim, B. Aksanli, A. Kumar, and T. S. Rosing, "Hierarchical and distributed machine learning inference beyond the edge," in *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, May 2019, pp. 18–23.

[20] H. N. Saha, S. Tapadar, S. Ray, S. K. Chatterjee, and S. Saha, "A machine learning based approach for hand gesture recognition using distinctive feature extraction," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan 2018, pp. 91–98.