Interactive Visualization for Topic Model Curation

Guoray Cai
Penn State University
University Park, PA, USA
cai@ist.psu.edu

Feng Sun
Penn State University
University Park, PA, USA
fzs122@psu.edu

Yongzhong Sha
Lanzhou University
Lanzhou, Gansu, China
shayzh@lzu.edu.cn

ABSTRACT

Understanding the content of a large text corpus can be assisted by topic modeling methods, but the discovered topics often do not make clear sense to human analysts. Interactive topic modeling addresses such problems by allowing a human to steer the topic model curation process (generate, interpret, diagnose, and refine). However, human have limited ability to work with the artifacts of computational topic models since they are difficult to interpret and harvest. This paper explores the nature of such challenges and provides a visual analytic solution in the context of supporting political scientists to understand the thematic content of online petition data. We use interactive topic modeling of the White House online petition data as a lens to bring up key points of discussions and to highlight the unsolved problems as well as potentials utilities of visual analytics methods.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces: -*visual analytics*; H.4.2 Information Systems:: -*visual analytic systems*

Author Keywords

Topic models; Information visualization; visual analytics

INTRODUCTION

Topic modeling has been advanced as a solution to the challenge of making sense of large corpora of textual data. With the help of machines, valuable themes buried in a large document collection can emerge and provide a better representation of the documents. The most popular topic modeling techniques, LDA (Latent Dirichlet Allocation) [4] and its variants, such as supervised LDA [26] and supervised anchor LDA [3], have been proven useful in many applications [29, 25], including online petition analysis [14]. Topic modeling assists qualitative and quantitative research over user-generated texts coming from the blogs or social media. By studying the set of topics learned from social media conversations over some period of time, it may become possible to find out what users are talking about, identify underlying topical trends, and follow them through time. Topic similarities among documents also help to identify the most relevant documents for a specific topic. Ideally, an analyst may be able to draw conclusions from word distributions for topics and use such insight to conduct a more in-depth study on documents with high affinities for specific topics.

Despite such advances, topic models have not been widely adopted by data analysts for practical use of understanding large corpora [23]. Topics discovered by LDA and other algorithms often have both "good" and "bad" topics judged by users. Topics could be bad because (1) they often confuse two or more themes into one topic; (2) they often pick up two different topics that are (nearly) duplicates for human; and (3) nonsense topics [18], (4) topics with too many generic words (e.g., "people, like, mr") [5], (5) topics with disparate or poorly connected words [22], (6) topics misaligned with human interpretation [9], (7) irrelevant topics [27], (8) missing associations between topics and documents [11], and (9) multiple similar topics [5]. The presence of poor-quality topics has been cited as the primary obstacle to the acceptance of statistical topic models outside of the machine learning community [22]. The root of these problems lies in the fact that the objective function that topic models optimize does not always correlate well with human judgments of topic quality [7]. Due to these problems, the use of topic models to analyze domain-specific texts often requires manual validation of the latent topics to ensure that they are meaningful [16].

Addressing the above issues to make topic models usable by analysts who are not machine learning experts, a variety of human-in-the-loop methods have been proposed to allow analysts to manipulate and incrementally refine a topic model of a target text corpus [17, 18, 19, 2]. These methods typically involve the use of interactive visualization and direct manipulation of topic models to diagnose poor topics and fix them through operations such as adding or removing words in topics, adjusting the weights of words within topics, splitting generic topics, and merging similar topics [17]. For example, ITM [18] allows users to add, emphasize, and ignore words within topics, while UTOPIAN [8] allows users to adjust the weights of words within topics, merge and split topics, and create new topics. Additionally, iVisClustering [19] lets users manually create or remove topics, merge or split topics, and reassign documents to another topic, with the help of visually exploring topic-document associations in a scatter plot.

While these operations can be supported by direct manipulation and algorithmic extensions, it is more challenging to diagnose the quality concerns of machine-discovered topics, and in assessing if a refinement strategy results in topic im-

ESIDA '18, March 11, 2018, Tokyo, Japan

provement. This is where interactive visualization methods are most helpful. Topic Browser [6] uses a tabular visualization technique to assist assessing term orders within each topic, and Termite [10] focuses on supporting effective evaluation of term distributions associated with LDA topics through visualizations. TopicNets [13] used a web-based interactive visual interface to enables users to discover topics of increasing granularity through an informed selection of relevant subsets of documents.

While these visualization tools help users to assess and refine static topic models, they run short in supporting the whole topic curation process. *Topic model curation* goes beyond human validation of machine-generated topics to include the whole human-directed process of discovering topics that are useful specific to a domain of applications. For example, public opinion researchers may be interested in discovering what is the range of policy preferences expressed in blog-spheres. Crisis managers may be interested in conversations in social media that are especially informative to their decisions on how to allocate resources and dispatch rescue teams. For such applications, the use of topic models is not a one-shot process but is a broader process of seeking, assessing, relating, and structuring topics with the help of supervised and unsupervised topic models. A typical topic curation process starts with a vanilla topic model (purely unsupervised probabilistic model such as LDA), and let users conduct a full diagnostics to recognize good and bad topics. Good topics will be collected and kept in a "bag", while bad topics improved or removed. For the set of bad topics, users may explore multiple ways to adjust topic models (merging/splitting topics, adding/removing words from a topic, modifying orders or weights of words in a topic). Depending on the consequence of imposed correlations and constraints, a new round of modeling and refinement can be initiated to explore the topic space of the document collection either in breadth or depth.

Towards supporting topic curation, this paper focuses on understanding the specific challenges of topic curation in the context of analyzing online petition data. We gained insight by actually practicing interactive topic modeling on the petition data we collected from the White House online petition website "We the People". This data set is considered a unique source for understanding citizens' policy concerns and preferences [15]. The insight gained from this practice is used to inform the design of a visual analytic system that supports topic model diagnostics, refinement, and evaluation. We reflect the use of visual analytic methods to enable users to interactively curate topic models.

INTERACTIVE TOPIC MODELING OF PETITION DATA

Electronic petitioning (e-petitioning) is becoming a prevalent form of political action for enabling direct democratic engagement [20]. The data used for this study comes from the online petitioning platform "We the People", hosted by the White House. It contains 5,177 petitions accumulated over the course of six years (2011-2016). We further selected 4,095 petitions that are in English. Each petition has four fields: (1) a petition ID, (2) a title, (3) a description, and (4) category tags.

As topic models treat documents as "bag-of-words", the first step of preparation before model training is tokenization, which splits each petition into a set of words. As words may have various forms, lemmatization is then applied to transform them into a common base form. Compared with the stemming technique that shares a similar goal, lemmatization takes advantage of vocabulary analysis and thus can produce the dictionary form of words that users can interpret. Bigrams are also used here for performance purpose [32]. Finally, stopwords are removed from the texts, as well as the overly common terms that appear frequently (top 50), to avoid possible discrimination. The resulting corpus contains 11,189 unique terms.

System Design

Figure 1 shows the user interface of interacting with petition documents and topic words. This system has two functional areas. The lower part is a topic-word visualization that supports direct manipulation of words-to-topics correlation.

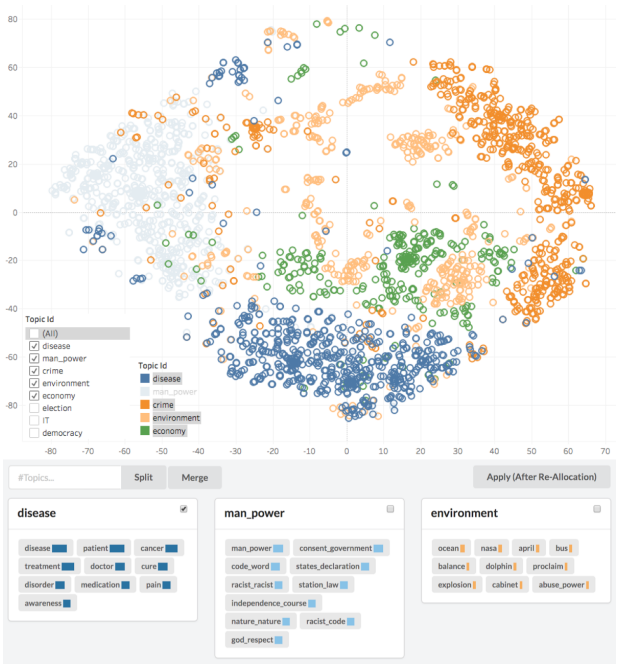


Figure 1: User interface for interactive topic modeling during exploration of petitions

The upper part is designed for exploring the topic quality from the perspective of how the petitions (documents) are clustered according to the space defined by the topics. The points cloud map provides a visual overview of the petition space where topically similar petitions are positioned adjacently. It is generated using t-SNE (t-distributed stochastic neighbor embedding) [8] to reduce the high-dimensional petition data to a 2-D vector space that human can perceive easily. Due to its nature of being nondeterministic, t-SNE usually transforms a high-dimensional data point to a different 2-D vector. However, the relationships between the data points will remain almost the same. An example of visualized petitions is shown in the Figure 1.

Each petition is assigned to one cluster based on its most salient topic and is color-coded correspondingly. Users can apply filters and highlighters on topics to manipulate the petition overview map. Highlighting enables users to review petitions in context while filtering allows users to focus on the petitions of interest. When hovering over a document point, a pop-up window displays the title, body, and topics of the document. In the meantime, the topic distribution (in terms of weights) of the selected document is visualized as a bar chart. By clicking a topic label, its topic-words distribution is visualized as color-coded bars.

At the back-end of the system, we choose Correlation Explanation (CorEx) [30] as the topic modeling algorithm to perform interactive topic curation. Built on the theory of Correlation Explanation [31] in information science, CorEx strives to represent the substrate information in a document collection that maximizes the informativeness of the data. Due to its fast training time and capability of supporting anchoring, CorEx can be easily tailored to incorporate human imposed correlations or constraints for semi-supervised topic modeling, making it an ideal choice for supporting interactive topic modeling [12]. Using CorEx, users can anchor multiple words to one topic, anchor one word to multiple topics, or any other creative combination of anchors in order to discover topics that do not naturally emerge. By leveraging CorEx's capability of topic seeding through *anchor words* in our system, human analysts can incorporate their knowledge and insights into the process of refining topic models.

TOPIC CURATION

Using our system for topic curation involves three phased of activities, with a number of iterations.

Topic Discovery

The first step is to use topic modeling algorithm with random seeds to run an unsupervised discovery of topics. The user must specify how many topics is to be produced, with the understanding that different numbers of topics can be chosen to analyze the petitions data on different levels of granularity and it is likely to generate a different set of topics [14, 24].

After initial unsupervised topic modeling with CorEx, users assess the topic model and conduct diagnostic analysis on topics. In particular, users will inspect topics, both individually and as a group, to evaluate their qualities by examining topic words. Those topics that users recognize as good ones should be kept. For those bad ones, users can file complaints and come up with one or more strategies to address them.

Topic Refinement

Topic refinement is achieved through manipulating topics-word representations at the bottom part of Figure 1. We included an anchoring mechanism to be coupled with CorEx models. It allows users to anchor one or more words to one topic, anchor one word to multiple topics, and anchor one or more words to some topics while not others. With this anchoring mechanism, topic revision interactions are supported by operations such as *splitting a topic*, *merging by joining*, and *merging by absorbing* (following [17]). More complicated operations can be achieved through a combination of above basic operations. For example, investigating more fine-grained topics can be accomplished by splitting topics iteratively.

Split a topic

If a topic is considered be "bad" based on the observation that it confuses two or more meaningful topics into one topic, a solution could be to split the topic into two or more topics. To do so, the user can check the topic he/she intends to split and then click the "split" button. Before applying the operation, the user is provided with the option to configure the number of resulting topics. Once confirming, the underlying model training will re-run under the new constraint that only the selected topic is decomposed while the others remain the same in terms of word allocation. Updated results will be generated and visualized.

In the backend, splitting a topic into n topics involves training a *word2vec* model to produce word embeddings [21]. The resulting model is used to calculate the semantical similarity between words. After that, a similarity matrix of the words within this topic is produced, and spectral clustering is applied to the matrix to categorize the words into n clusters. The n clusters of words are encoded into the previous model as anchor words and will produce n new topics to replace the original one.

Merge topics by joining

If several topics are judged to have something common in their semantic meaning, they can be merged into one topic. This is accomplished by selecting these topics and then clicking "apply" button. The system automatically apply the constraint that words assigned to the topics to be merged have to appear in the resulting topic. underlying model will be updated. Accordingly, the visualization will be re-rendered. In the backend, the words that appeared in the two topics are now anchored under the same one.

Merge topics by absorption

If one or more words in a topic are considered intruders and fit better to a different topic, the user can re-allocate topic words through drag-and-drop operations. Specifically, a user can select a word that is considered allocated incorrectly and move it to a more related topic. After reallocation of words is done, the petition view will update to reflect the modification. In the back end, *Merging topics by absorbing* is basically a reallocation process where selected words in one topic is anchored to the other one and a new model is trained. The rest of the topic-word assignments remain the same through anchoring as well.

Evaluating Topics Interactively

Evaluating the quality of the topics in the current model is necessary for both the diagnoses of good/bad topics as well as assessing the impact of topic revisions. Evaluating topic quality is done by assessing two aspects: (1) *are the words in a topic coherent and contributing to some collective meaning*? (2) *are the topics aligned with the information needs of the intended application*? As such, we designed the interface in Figure 1 that visualizes topically represented petitions to

support the following functions for evaluating the quality of topics:

Inspecting quality of every single topic. Users can evaluate topics by looking at the coherence of the component words and their relative weights (see the bars next to words) on a topic. Topics are also color-coded in the visualization window. Clicking on the legend of a topic results in all the petitions with sufficient weights on that topic being highlighted (while other petitions are dimmed). These functions allow users to explore the patterns of how petitions of the same topic clustered. A good topic tends to create a cluster of petitions that are less mixed with petitions.

Comparing topics. Users can evaluate one or more topics together by observing semantic relations to spatially close or remote topics, and by looking at the spatial relationships (overlapping clusters, adjacent clusters, non-intersecting clusters) between petitions of the two topics. Applying filters to leave fewer topics on the figure helps reduce visual clutters.

TOPIC MODEL CURATION SCENARIO

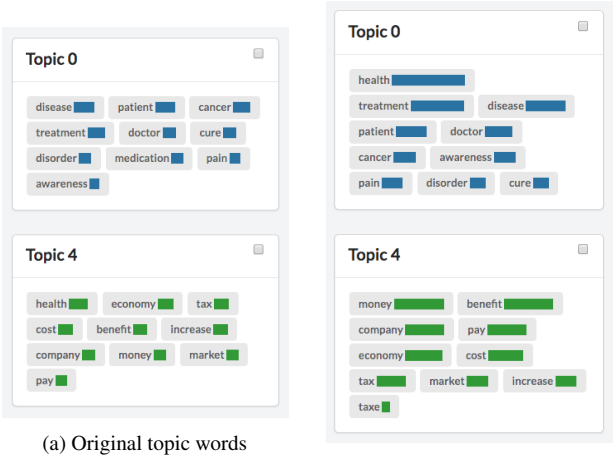
We practiced topic curation process on the online petition dataset to experience how well our system supports topic diagnostics and refinement. Firstly, we run the CorEx topic modeling and generated 20 topics. A fixed random seed was used to make sure the same results can be reproduced. Table 1 shows 5 samples out of 20 produced from a topic model. The initial result from the CorEx topic modeling reveals interesting topic clusters from the data set. In the provided samples, topic 0 mainly talks about "disease", topic 4 generally discusses "economy", topic 5 describes "election", and topic 16 represents "law enforcement". The bottom part of the table shows the results after applying certain topic revision operations.

Table 1: Selected topics (#topics = 20)

id	topic words (top 15)
0	disease, patient, cancer, treatment, doctor, cure, disorder, medication, pain, awareness, symptom, illness, medicine, diagnosis, disability
4	health, economy, tax, cost, benefit, increase, company, money, market, pay, healthcare, fund, research, dollar, debt
5	election, investigation, vote, voter, candidate, hillary_clinton, voting, campaign, department_justice, fbi, ballot, office, corruption, violation, democrat
6	internet, consumer, energy, information, technology, provider, service, device, car, access, fuel, safety, standard, road, vehicle
16	officer, police, law_enforcement, evidence, police_officer, county, aircraft, judge, governor_chris, killing, conviction, department, scene, cat, chief
0'	health, treatment, disease, condition, patient, doctor, cancer, awareness, pain, illness, medicine, disability, disorder, cure, medication
4'	money, benefit, company, pay, economy, business, cost, fund, tax, industry, dollar, budget, study, market, increase
6.1	service, information, com, access, standard, technology, internet, consumer, provider, content, http, privacy, https_facebook, internet_service, customer
6.2	safety, vehicle, energy, car, device, accident, fuel, road, aviation, forest, traffic, emission, faa, air, carbon
5+16	investigation, vote, election, officer, police, law_enforcement, campaign, candidate, corruption, voter

Moving Intruder Words

By examining the above table, we find that topic 4 contains a word "health" that is clearly different from other words (see Figure 2). We also find that some petitions related to health but has nothing to do with "economy" are assigned to this topic during the petition exploration phase. One example petition is "place mental health as a required course in junior high and middle schools". In order to correct this topic assignment, we performed topic refinement by moving the intruder word "health" from topic 4 to topic 0. The re-generated topic words are shown in Table 1 as topic 0' and topic 4'.



(a) Original topic words

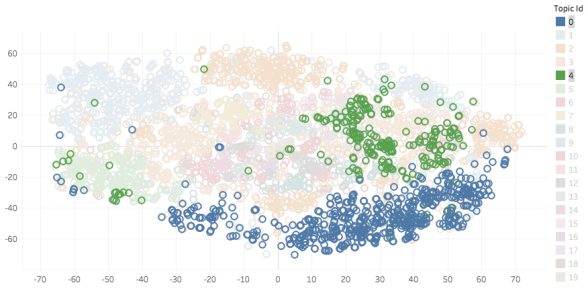
(b) New topic words

Figure 2: Move topic word "health" from topic 4 to topic 0

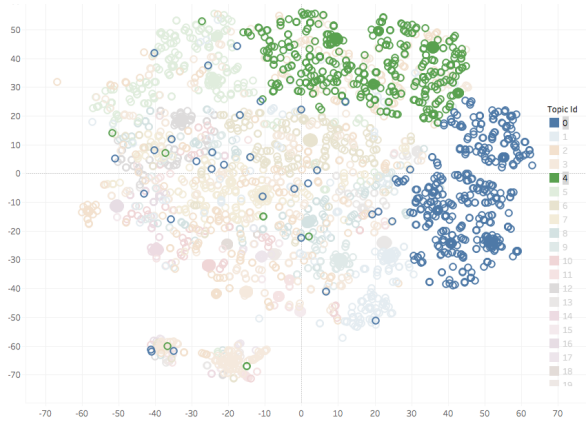
In order to assess if such a strategy of refining topics has led to a better outcome, we rendered the petition clusters in relation to the new topic definition and the result is shown in Figure 3. From this figure, we can clearly see how topic groups are isolated and cut. Compared with Figure 1, outliers are nicely scattered apart and small clusters of outliers disappear. Such result suggests that the change of topic model by moving "health" from topic 4 to topic 0 is a good move. This claim is further confirmed by a calculated metric of topic coherence based on word context vectors [1]. This metric has been demonstrated to have the highest correlation with the interpretability of topics [28]. The topic coherence of topic 4 is increased from 0.453 to 0.555 after removing the word intruder, and the overall topic coherence is increased from 0.431 to 0.443.

Split a Multi-theme Topic

Observations show that the distribution of petitions of topic 6 is scattered in the reduced-dimensional space: there are several small clusters of petitions. By sampling some of them for detailed inspection of petition contents, we found that some semantically irrelevant petitions are placed adjacently in the visualization, e.g., "Prevent the FCC from ruining the Internet" and "Put a fee on carbon-based fuels and return revenue to households", the former is about Internet and information technology, while the latter is related to energy. This finding can also be validated by examining topic words of topic 6: "internet", "information", and "technology" are clearly incoherent



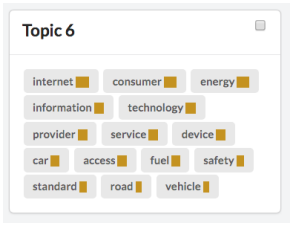
(a) Petitions of topic 0 and topic 4



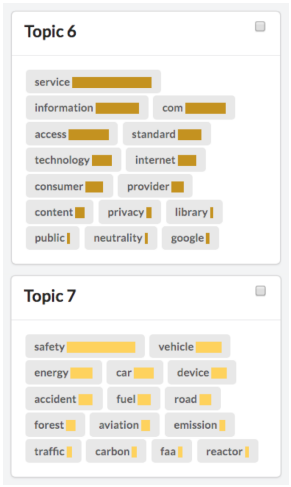
(b) Petitions of topic 0' and topic 4'

Figure 3: A comparison of visualized petitions before and after moving words between topic 0 and topic 4

with "energy", "fuel", and "safety". Therefore, we believe topic 6 is of low quality since it contains several sub-topics and needs to be diluted.



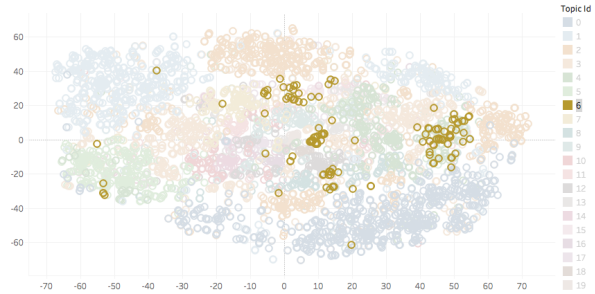
(a) Original topic words



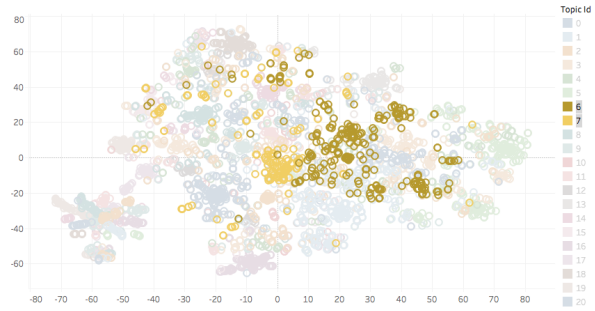
(b) New topic words

Figure 4: Split topic 6 into two topics topic 6 (6.1) and topic 7 (6.2)

To address the quality concerns of topic 6, we split topic 6 into two topics (by clicking on Topic 6 and choose "Split" button).



(a) Petitions of topic 6



(b) Petitions of topic 6 (6.1) and topic 7 (6.2)

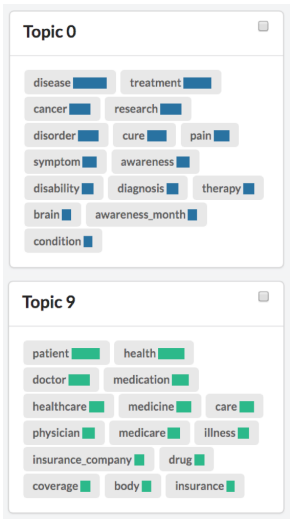
Figure 5: A comparison of visualized petitions before and after splitting topic 6

The modified version of the topic model is shown in Table 1 as 6-1 and 6-2 and Figure 4 as 6 and 7. The figure shows that the weights of the first several topic words are increased, indicating that these words can better represent the topics. It is also apparent from Figure 5 that the distributions of petitions for topic 6 and topic 7 become more focused, indicating that the petitions documents within same clusters are more topically homogeneous. After the new topic model applied, the above example petitions are allocated to the correct topics respectively, resulting in an increase of overall coherence value from 0.431 to 0.441. Specifically, the original topic 6 has an individual coherence score of 0.341, while the scores of newly produced topic 6 and topic 7 are 0.594 and 0.419 respectively.

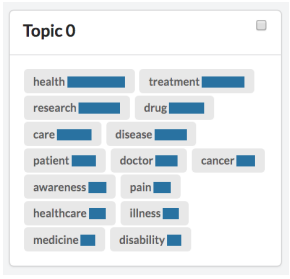
Merge Semantically Similar Topics

If the number of topics is set to a large number, CorEX algorithm will generate topics in finer granularity of topics. This could create situations where words that contribute to a single theme end up in separate topics. Under such circumstance, a merging operation is necessary to make sure that petitions of similar topics are grouped together. In order to demonstrate this situation, we trained another topic model by setting the number of topics as 50 (relatively large) and the topic words are shown in Figure 6a. By looking at the topic words, topic 1 and topic 7 both describe "healthcare" but appear to be different topics.

The topic words after merging these two topics are shown in Figure 6b. Petitions of these two topics are now grouped into one cluster as well. Subsequently, these petitions can be processed and analyzed as a whole, e.g., summarized and forwarded to the Department of Health and Human Services.

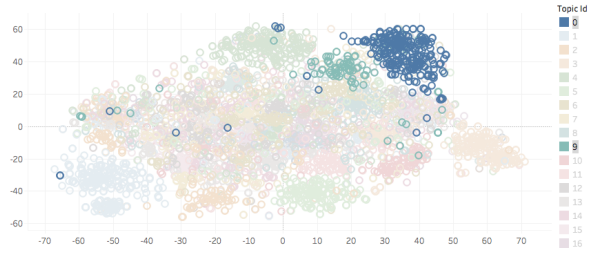


(a) Original topic words

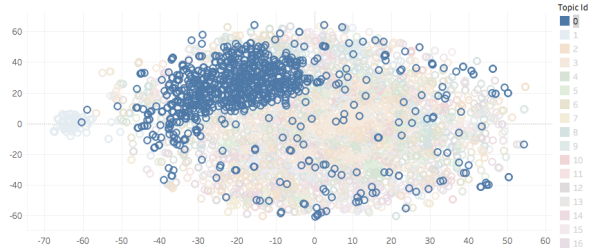


(b) New topic words

Figure 6: Merging topic 0 and topic 9



(a) Petitions of topic 0 and topic 9



(b) Petitions of topic 9 (0 + 9)

Figure 7: A comparison of visualized petitions before and after merging topic 9

Merging topics is also useful when a small number of topics is used. Referring to the before-mentioned topic model of 20 topics, we found that topic 5 contains words “investigation” and “justice” that may be related to topic 16. Therefore, we performed a merging by joining on these two topics and it leads to a more general topic denoted as 5+16. Although the coherence value of merging the two topics remains almost the same, it is noteworthy that a new word “corruption” is prioritized as it could serve as a bridge to connect two topics represented as “election” and “law enforcement” (e.g., a petition titled “Arrest and prosecute officials who tried to suppress the vote in the 2012 election”), showing that merging topics has the potential of revealing latent relationship among them.

Topics that are difficult to interpret may still exist even after several iterations of topic refinements. On the other hand, some petitions are complicated in that they have multiple equally important aspects and even people have difficulty in identifying the most representative one. For those documents that are related to "bad" topics and can not be fixed at this round of analysis, the system can collect them into a subset of data to be fed into the next round of analysis.

DISCUSSION

Our work on analyzing the topic structures of online petitions is still a work-in-progress, but we have gained several lessons about interacting with topic modeling tools. First, users have to deal with tremendous uncertainties when deciding what is the proper strategy in tuning the topic model. Visualizing the impact of multiple strategies and providing interaction capabilities to assess the quality of topics and compare the document clusters before and after the model tuning will be critically important.

Another finding from this exercise is that there is a need to construct topic hierarchy from unsupervised topic models in order to be aligned with the way political scientists perceive the world of petition data. However, the topics discovered by CorEx algorithm have a flat structure, and they tend to be biased towards those topic branches that have more detailed data. We will continue to explore our visual analytic approach for incremental refinement of topic structures and demonstrated how such an approach can be used to uncover topic hierarchy of petitions that best reflects the human conception of the domain. Further work is required to evaluate the usability and effectiveness of this method. While we used dimension reduction based visualization, other petition explorations and analysis approaches should be investigated as well.

ACKNOWLEDGEMENT

The authors would like to acknowledge funding support from National Science Foundation under award # IIS-1211059, and from a grant funded by the Chinese Natural Science Foundation under award 71373108.

REFERENCES

1. Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. 13–22.

2. David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML ’09)*. ACM, New York, NY, USA, 25–32.

3. Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*. 280–288.

4. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
5. Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Its Applications*. Chapman & Hall, Chapter 12, 225 – 254.
6. Ajb Chaney and Dm Blei. 2012. Visualizing Topic Models.. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. 419–422.
7. J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems.* 288–296.
8. Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEE Transactions of Visualization and Computer Graphics* 19, 12 (2013), 1992–2001.
9. Jason Chuang, Sonal Gupta, Christopher D Manning, and Jeffrey Heer. 2013. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. In *Proceedings of the 30th International Conference on Machine Learning*. 612–620.
10. Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite : Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*. 74.
11. Hal Daumé. 2009. Markov random topic fields. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* August (2009), 293–296.
12. Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2016. Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *arXiv preprint arXiv:1611.10277* (2016).
13. Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Trans. Intell. Syst. Technol.* 3, 2, Article 23 (Feb. 2012), 26 pages.
14. Loni Hagen, Ozlem Uzuner, Christopher Kotfila, Teresa M. Harrison, and Dan Lamanna. 2015. Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach. In *2015 48th Hawaii International Conference on System Sciences*, Vol. 2015-March. IEEE, 2134–2143.
15. Scott A Hale, Helen Margetts, and Taha Yasseri. 2013. Petition growth and success rates on the UK No. 10 Downing Street website. In *Proceedings of the 5th annual ACM web science conference*. ACM, 132–138.
16. David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 363–371.
17. Enamul Hoque and Giuseppe Carenini. 2016. Interactive Topic Modeling for Exploring Asynchronous Online Conversations. *ACM Transactions on Interactive Intelligent Systems* 6, 1 (feb 2016), 1–24.
18. Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning* 95, 3 (2014), 423–469.
19. Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum* 31, 3pt3 (2012), 1155–1164.
20. Ralf Lindner and Ulrich Riehm. 2009. Electronic petitions and institutional modernization. International parliamentary e-petition systems in comparative perspective. *JeDEM-eJournal of eDemocracy and Open Government* 1, 1 (2009), 1–11.
21. Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013).
22. David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 2 (2011), 262–272.
23. Sergey I. Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43, 1 (2017), 88–102.
24. Paul Hitlin. 2016. *'We the People': Five Years of Online Petitions*. Technical Report. Pew Research Center.
25. Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing Microblogs with Topic Models. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010), 1–8.
26. Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009a. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 248–256.
27. Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009b. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* August (2009), 248–256.

28. Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 399–408.

29. Amin Sorkhei, Kalle Ilves, and Dorota Glowacka. 2017. Exploring Scientific Literature Search Through Topic Models. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics (ESIDA '17)*. ACM, 65–68.

30. Greg Ver Steeg and Aram Galstyan. 2014. Discovering Structure in High-Dimensional Data Through Correlation Explanation. In *Advances in Neural Information Processing Systems, NIPS'14*.

31. Greg Ver Steeg and Aram Galstyan. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*. 577–585.

32. Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 90–94.