Bringing Insights into a Corpus: Topic Trends Detection and Labeling

Heidar Davoudi

School of Computer Science and Engineering
York University
Ontario, Canada

Abstract. Probabilistic topic models based on Latent Dirichlet Allocation (LDA) are increasingly used to discover hidden structure behind big text corpora. Although topic models are extremely useful tools for exploring and summarizing large text collections, the inferred topics are not easy to understand and interpret for human. This project aims to provide more insights into an available corpus using domain knowledge (i.e., domain glossary) provided by experts. We proposed a similarity measure based on the rank of words in inferred topics and normalized pointwise mutual information to measure the coherence between the topics and domains. The labeled topics are considered over time to evaluate the dynamics of the corpus.

Keywords: Topic modeling, topic labeling, topic trends

1 Introduction

Probabilistic topic models are increasingly used for unsupervised analysis of big corpora. While different models (e.g., pLSI [1] and mixture of unigrams) were proposed for discovery of hidden structure of text data, Latent Dirichlet Allocation (LDA) [2] drew more attentions in the computational linguistic community due to its realistic underlying assumption. The basic idea of LDA is to capture latent semantics in a big corpus even though it can be utilized in other domains as well. For example, in social science text analytics are widely used as a way of unobtrusively observing people and their interactions, where words are treated as a proxy of secondary phenomenon [3].

Despite the fact that topic models are quite useful algorithmic tools to explore and summarize the corpus, whether the latent space is interpretable by human needs to be evaluated. One possible solution is to present topic distributions and document-topic assignments to users, and then assess the users' judgments. For example, Change et al. [4] proposed a quantitative method to evaluate semantic meaning in inferred topics. They designed two user-centric evaluation tasks aimed at evaluating the quality of both inferred topics as well as the topic to document assignments. However, the proposed method used human judgments to examine the topics.

In another research, Newman et. al [5] evaluated the coherence of inferred topics by applying different scoring measures and using WorldNet, Wikipedia and Google search engine as the external sources. For Wikipedia and WordNet, the pairwise scores between the words of two topics with respect to the external source were used as the coherence measurement. In case of Google, they used search engine-based similarities (i.e., Google title matches and Google log hits matches) as the topic coherence criteria.

This projects aims to provide more insights into the available corpus by answering some questions: to what degree are the knowledge provide by experts (domain glossary) and the corpus related to each other?, can the topics extracted from the corpus be labeled consistently?, to what extend are the topics correlated to each other over time? and to what degree are correlations between the topics over time consistent with the topic labels.

This project organized as follows: section 2 introduces the topic models and compares LDA to other probabilistic models. Section 3 deals with the topic model construction, and section 4 elaborates the proposed topic labeling method. The experimental results and discussions are presented in section 5, and finally we have the conclusion in section 6.

2 Building a Topic Model

The basic idea behind topic models is that documents are mixtures of topics and each topic is a distribution over words. In fact, it is assumed that the topic model is a generative model or in other words, there is an underlying probabilistic distribution producing documents.

2.1 Generative Process

Making a new document in LDA needs following probabilistic procedure known as the generative process for each document: choosing a distribution over topics, then assigning a topic to each word in the document and finally draw a word from that topic. The generative process for LDA can be described as follows:

1. For each topic $1..k$
   (a) Draw a multinomial over words $\beta_k \sim Dir(\eta)$
2. For each document $1...k$
   (a) Draw a multinomial over topics $\theta_d \sim Dir(\alpha)$
   (b) For each N word $w_{d,n}$
      i. Draw a topic $z_{d,n} \sim Multi(\theta_d)$
      ii. Draw a word $w_{d,n} \sim Multi(\beta_{z_{d,n}})$ from that topic

In fact, LDA (or other topic models) learning can be seen as the method of inverting this process and inferring the set of topics which generated the collection of documents. Therefore, the results of LDA are a distribution over words (for each topic) and a distribution over topics (for each document).

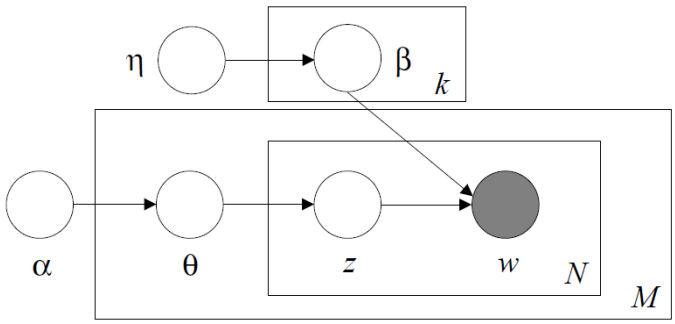

Fig. 1. LDA graphical model (adopted from [2]).

Transferring from word-based to topic-based representation of documents has some major advantages: topic models not only reduce the space dimensions intuitively but also provide a great insight into the corpus and each individual document. As a matter of fact, the topic modeling provides a structure in which the topics and the documents are individually interpretable. This results in useful consequences and applications in many domains (e.g., recommendation system).

2.2 Graphical Model

LDA generative process can be illustrated alternatively by a graphical model. The graphical model provides a simple way to visualize the underlying probabilistic assumption as well as the generative process needed to produce each sample. In topic modeling context, it is a directed graphical model comprising nodes (random variables) connected by links (probabilistic dependencies between random variables). The observed variables are denoted by shading the corresponding nodes; all other variables are known as latent variables. Plate notation allows more compact representation by surrounding a variable in a box, called plate, indexed by N indicating that there are N similar nodes. Figure 1 shows the graphical model representation of LDA using plate notation.

2.3 LDA Geometric Interpretation

Figure 2 shows the Dirichlet distribution for three topics in a two dimensional simplex. For any point in the simplex, sum of components equals to 1. This means that each point can represent a probability distribution whose parameters equal to coordinates of the point. Dirichlet prior to topic can be interpreted as a probability mass spread over the simplex. For $\alpha < 1$, the probability mass is located at the corners of simplex (there is a bias towards the sparse topics).

Figure 3 illustrates the simplex where the number of topics is two. Two bold points are the topics and the dot line represents the documents generated by the generative model. Each point of the dot line segment is a convex combination of the two topics. Interestingly, the hyper parameter η can be interpreted as a force

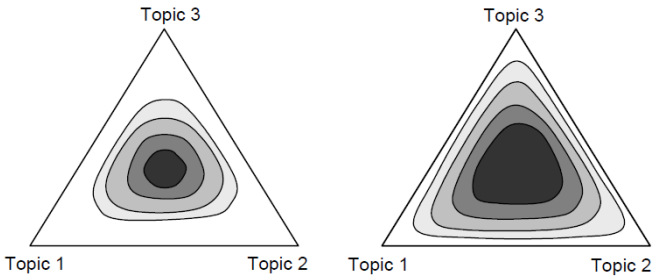

Fig. 2. Symmetric Dirichlet distribution in a two dimensional simplex (left: $\alpha = 4$, right: $\alpha = 2$), adopted from [6].

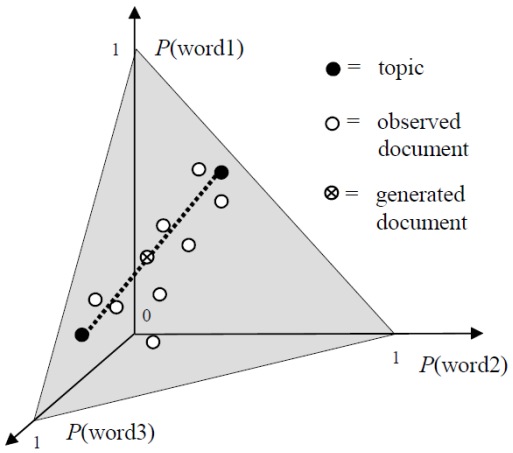

Fig. 3. A geometric Interpretation of LDA (adopted from [6]).

on topic locations while the hyper parameter α can be interpreted as a force on the document locations on the dot line segment.

2.4 LDA Versus Other Probabilistic Models

Figure 4(a) shows the other probabilistic model called unigram. In this model, each word is drawn independently from a single multinomial distribution. This model completely ignores the concept of topic in the corpus. Another model is mixture of unigrams which is illustrated in Figure 4(b), this model assumes that each document is generated by first choosing a topic (z) and then generating N words independently from that topic. Obviously, this model assumes one topic per document which is quite unreasonable. Probabilistic latent semantic indexing (pLSI) is the other topic modeling shown in Figure 4(c). The basic assumption in this model is that document d and word w_n are conditionally independent given an observed topic. Although the model assumes that each document is

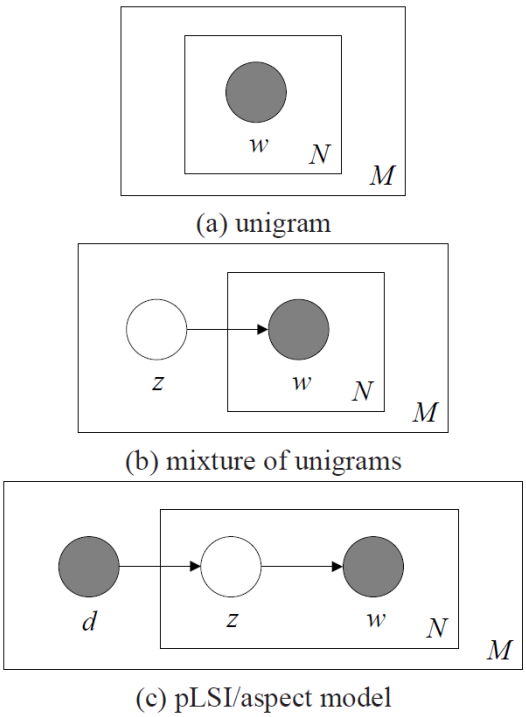

Fig. 4. Graphical model representation of unigram, mixture of unigrams, and pLSI.

generated by multiple topics ($p(z|d)$ can be seen as the mixture weight of topics for document d), serving d as a multinomial random variable (with many possible values), needs large number of documents in the learning process. Moreover, the large number of parameters in this model makes it prone to overfitting.

3 Inference and Parameter Estimation

Both inference and parameters estimation (using maximum likelihood method) need computing the posterior $p(\beta, \theta, z|\eta, \alpha)$. The following section describes the variation method to estimate the posterior distribution efficiently.

3.1 Variational Method

The main idea behind the vibrational method is to estimate the posterior probability $p(\beta, \theta, z|\eta, \alpha)$ by choosing a distribution like $q(\beta, \theta, z|\lambda, \gamma, \phi)$ from a family of distribution (e.g., exponential family). In mean field variation method, this family is characterized as follows:

$$q(\beta, \theta, z|\lambda, \gamma, \phi) = q(\beta|\lambda)q(\theta|\gamma)q(z|\phi) \tag{1}$$

The variational distribution has its own variational parameters: λ, γ *and* ϕ. In fact, the goal is to find the variational parameters that make variational distribution as close as possible to the posterior distribution. To measure the closeness of variational and posterior distribution, Kullback-Leibler (KL) divergence usually is used. So, the optimal value of the variational parameters can be found by setting the following optimization problem:

$$\arg\underset{\lambda,\gamma,\phi}{min} KL(p(\beta,\theta,z|\eta,\alpha) \parallel q(\beta,\theta,z|\lambda,\gamma,\phi)) \tag{2}$$

Although the variational distribution q with the tuned parameters can be used as a proxy of posterior distribution p, KL divergence usually cannot be minimized directly. However, KL divergence can be minimized indirectly by maximizing a lower bound drivable from Jensen inequality:

$$\log p(w|\alpha,\eta) \geq L(\lambda,\gamma,\phi|\alpha,\eta) \tag{3}$$

Moreover, it can be verified that:

$$\log p(w|\alpha,\eta) = L(\lambda,\gamma,\phi|\alpha,\eta) + KL(p(\beta,\theta,z|\eta,\alpha) \parallel q(\beta,\theta,z|\lambda,\gamma,\phi)) \tag{4}$$

Equations 3 and 4 are the base of variational EM algorithm for LDA (for each document) as follows:

1. **E-Step**: Find the maximum of $L(\lambda,\gamma,\phi|\alpha,\eta)$ with respect to λ,γ,ϕ.
2. **M-Step**: Find the maximum of $L(\lambda,\gamma,\phi|\alpha,\eta)$ with respect to α,η

In fact, in E-Step, we find the best q as the proxy of p (in the best case $p = q$, where the KL divergence equals to zero), and in M-Step, we find the parameters which decrease the $\log p(w|\alpha,\eta)$. More formally, it can be shown although we maximize the lower bound on likelihood, the likelihood is non-decreasing in each iteration:

$$\log p(w|\alpha^{k-1},\eta^{k-1}) \underset{E-Step}{=} L(\lambda^k,\gamma^k,\phi^k|\alpha^{k-1},\eta^{k-1}) \tag{5}$$

$$\underset{M-Step}{\leq} L(\lambda^k,\gamma^k,\phi^k|\alpha^k,\eta^k) \tag{6}$$

$$\underset{JensonInequality}{\leq} \log p(w|\alpha^k,\eta^k) \tag{7}$$

4 Topic Labeling

4.1 Pointwise Mutual Information (PMI)

The Pointwise Mutual Information (PMI) of two random variables x and y measures the difference between joint probability distribution of x and y, and their individual distributions. Mathematically, it can be defined as follow:

$$Pmi(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{8}$$

The measure is symmetric and can take both positive and negative values. PMI maximize when x and y are fully associated. The lower and higher bound of PMI value can be defined as follows:

$$-\infty \leq Pmi(x,y) \leq \min[-\log p(x), -\log p(y)] \tag{9}$$

4.2 Normalized PMI

The PMI measure can be normalized by dividing the equation by $-\log[p(x,y)]$:

$$Npmi(x,y) = \log \frac{pmi(x,y)}{-\log[p(x,y)]} \tag{10}$$

The resulting measure (i.e., Normalized Pointwise Mutual Information) range is between [-1,1], where -1 and +1 indicate never co-occurrence and fully co-occurrence. The independence between x and y can be quantified by the zero value.

4.3 Labeling Score

We utilized Normalized PMI defined in (10) to define the coherence score between the top ranked words (based on the probability of the word in the topic) in each topic and the vocabulary of each domain. Suppose that r_{it} is the rank of w_{it} (the $i'th$ word in topic t) and v_{jd} is the $j'th$ word in domain d, the coherence score between topic t and domain d can be defined as follows:

$$Score(t,d) = \sum_i \sum_j r_{it}\ Npmi(w_{it}, v_{jd}) \tag{11}$$

The topic label (domain) for topic t can be calculated as follows:

$$Label(t) = \arg\max_d Score(t,d) \tag{12}$$

While LDA explores the co-occurrence between the words in a particular topic, the proposed score tries to measure the co-occurrence between the words in a topic and the words in a particular domain.

5 Results and Discussions

5.1 Dataset

Table 1 shows the dataset used in this experiment. The dataset is a part of corpus introduced in [7]. Truth table consists of 20 documents which experts found them related to forced migration. Dataset 1 and 2 comprise two corpora collected in February and March, 2014 respectively. while dataset 1 contains much ore documents and collected over a month, dataset 2 includes only 35,552 documents and was collected in one day. The original dataset also includes a glossary for domains of interest where each domain is characterized by a set of words. Table 2 shows the domains as well as the number of terms related to each domain.

Table 1. Dataset used in the experiment.

Dataset	No. of Doc.	Time Period
ground truth	20	-
dataset 1	421,644	2014-02
dataset 2	35,552	2014-03

Table 2. Domain characteristics.

Domains	No of terms
Relief	57
Governance	105
Economic	49
Demographics/Identity	69
Environmental/Biological	104
Infrastructure	123
Violence	77
Holidays	12
Movement	17
Prominent Figures/Entities	148

5.2 Results

Table 3 shows the top-15 words of five topics inferred for dataset 1, where the number of topics equals to 10. While some topics like topics 4 and 6 seems to be coherent (topic 4 as "sport" and topic 6 as "business"), some topics (like topic 2) are quite meaningless. This shows that the corpus contains a lot of noises which need to be removed. However, capturing all noises in some specific topics is quit interesting and can be used in the corpus cleansing.

Figure 5 compares the perplexity of dataset 2 predicted by LDA models trained with ground truth and dataset 1. Interestingly, the perplexity is very low for the model trained with the ground truth dataset when the number of topics is 10. It could show that both ground truth and dataset 2 have very narrow topics and there are considerable overlaps between the contents of two datasets topics. As the number of topics increases, dataset 1 leads to the better perplexity due to the high number of training documents.

In order to find the relationship between the domains characterized in the glossary and the available corpus (dataset 1), the document frequency for each terms in the glossary is calculated. The result of document frequency analysis is illustrated in Figure 6 where the vertical axis shows the average number of documents containing related words in a particular domain. As it can be seen, "Infrastructure" and "Governance" domains are more frequent than the others while "holidays" and "Prominent Figures/Entities" domains are the least frequent ones. The "movement" domain has the medium vocabulary in the corpus.

Assuming a document is a mixture of topics, the number of documents containing a topic in a day is the sum of probabilities of the topic given documents.

Table 3. Top-15 words of five topics in dataset 1.

Topic 2	Topic 4	Topic 6	Topic 8	Topic 9
de	team	year	show	government
la	game	million	time	president
sa	games	percent	film	people
le	season	company	day	country
na	time	market	year	minister
ang	year	billion	people	security
des	league	cent	love	police
din	win	business	world	china
les	play	bank	music	ukraine
en	back	growth	life	february
ng	sochi	companies	story	state
au	players	price	years	military
cu	club	industry	article	foreign
pe	world	sales	die	political
care	olympic	data	make	russia

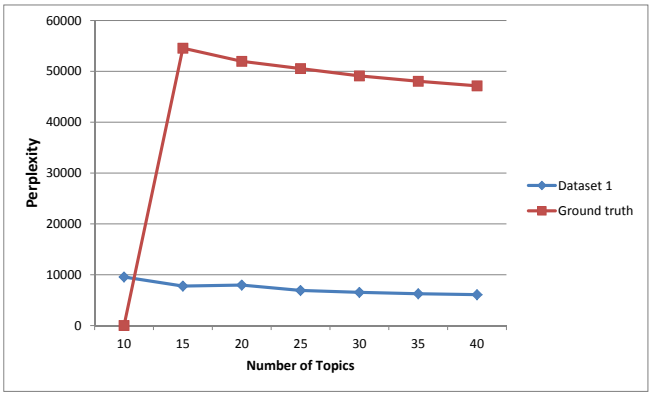

Fig. 5. Perplexity of LDA on dataset 2 (trained by ground truth vs. dataset 1).

Figure 7 shows the topic trends for dataset 1 (one month). The vertical axis shows the number of documents influenced by the inferred topics. The domain labels indicated in the graph are calculated using the method proposed in Section 4.3. As it can be seen, "Infrastructure" is the most influential topic for dataset 1 (it is assigned to three topics in comparison to "Violence", "Governance" and "Economic" which are assigned only to two topics). In fact, the labeling procedure produced 5 distinct topics. Moreover, by considering the topic trends over time, it can be seen that some labeling results are quit consistent with the trend correlations (e.g., Infrastructure (1) and (2), Governance (1) and (2)) even though some of them have different trends over time (e.g., Infrastructure(3), (2),

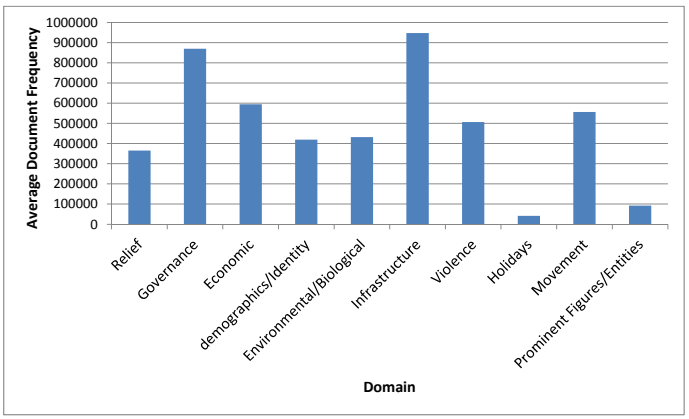

Fig. 6. Average document frequency for each domain.

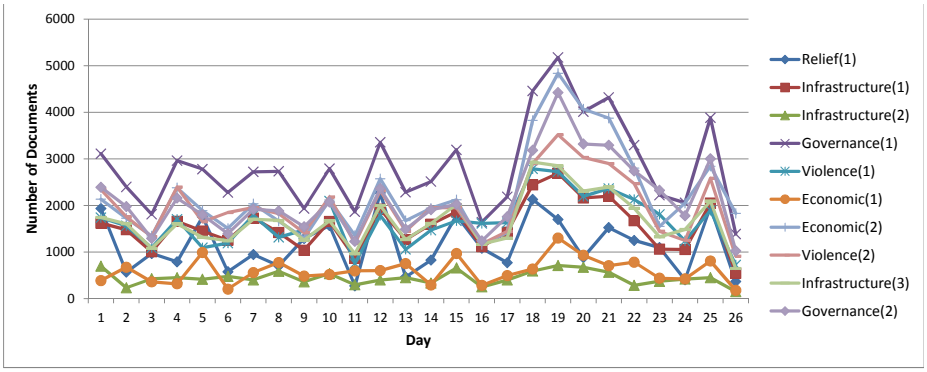

Fig. 7. Topic trends (dataset 1).

Economic(1) and (2)). Furthermore, there are some strong correlations between the domains labeled differently (e.g., Economic(2) and Governance (2)).

Figure 8 to 11 illustrate the topic trends over time based on the topic models inferred separately for each week in dataset 1. As it can be seen, the trained topics are more correlated on weekly data. For example, all topics in Figure 9 except Infrastructure (1) to (3) are highly correlated and topics in Figure 10 are divided into two highly correlated groups (Infrastructure (1), (3), (4), Relief(1) and Infrastructure(2), Governance (1), (2), (3) Violence (1), Economic(1)).

For all labeled topics in Figure 7 and Figure 8 to 11, the number of topics was 10. Figure 12, 13 and 14 illustrate the sensitivity of labeling procedure to number of topics for ground truth, dataset 1 and 2 respectively . As it can be seen, the labeling algorithm is not sensitive to number of topics since the proportion of

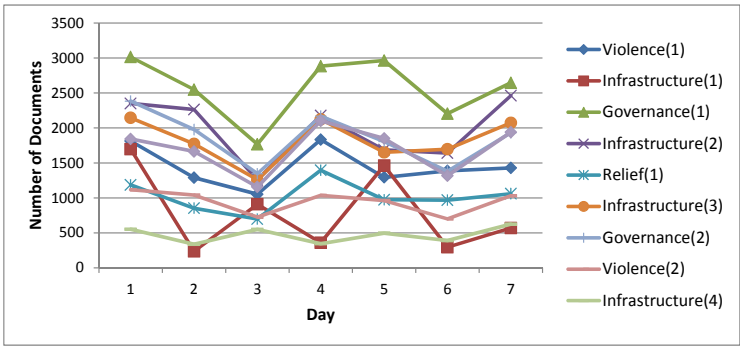

Fig. 8. Topics trends (dataset 1, week 1).

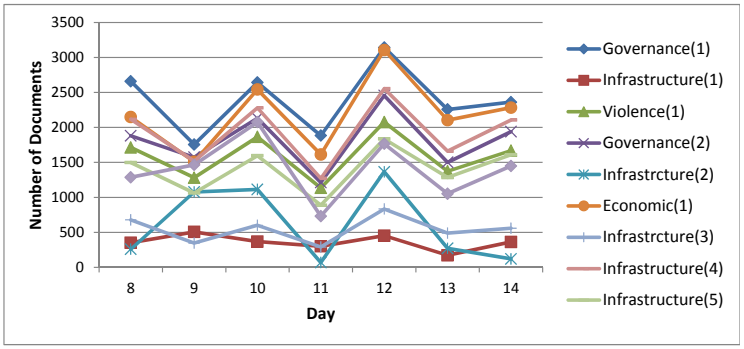

Fig. 9. Topics trends (dataset 1, week 2).

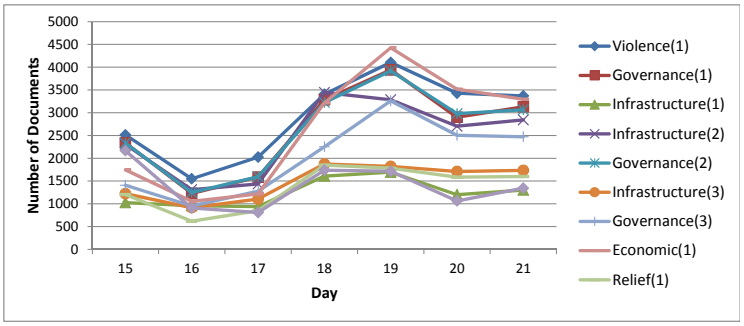

Fig. 10. Topics trends (dataset 1, week 3).

assigned labels to the topics does not change significantly as the number of topics increases.

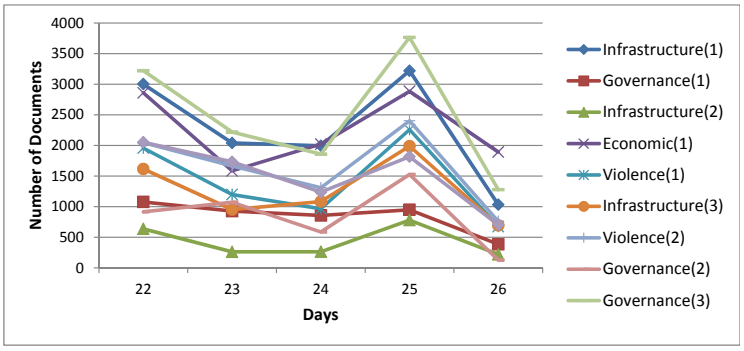

Fig. 11. Topics trends (dataset 1, week 4).

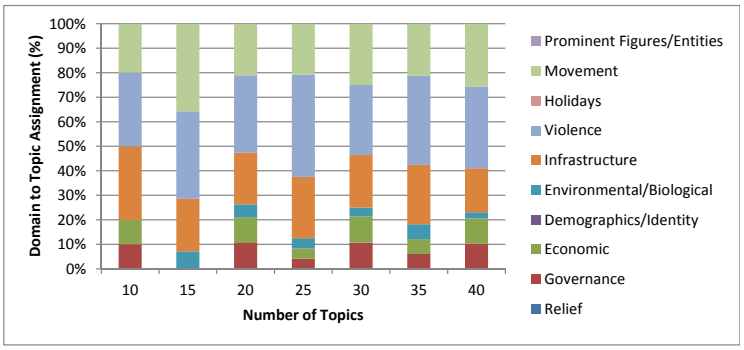

Fig. 12. Proportion of assigned domains (ground truth).

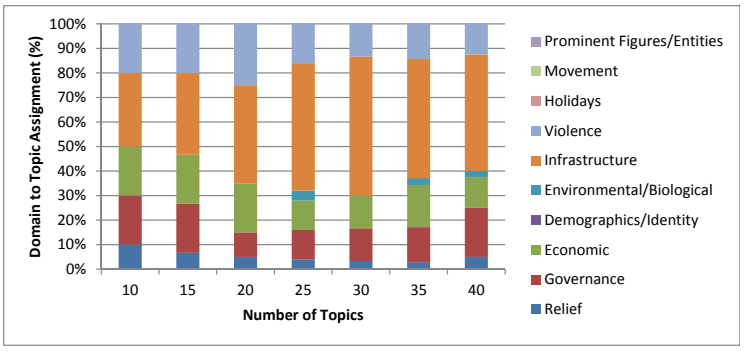

Fig. 13. Proportion of assigned domains (dataset 1).

6 Conclusions

In this project, we explored dynamic of a corpus with respect to an available domain knowledge. Latent Dirichlet Allocation (LDA) is used as an automatic

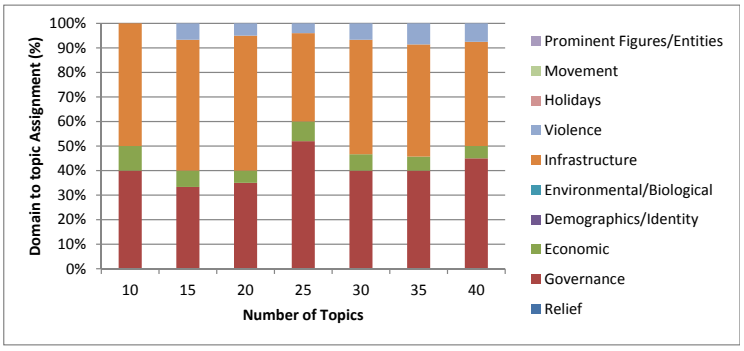

Fig. 14. Proportion of assigned domains (dataset 2).

algorithmic tool for detecting topics in the corpus. The inferred topics are labeled based on the devised similarity measure and then considered over time. The results showed that there were strong correlations among the topics derived from LDA. Moreover, the number of distinct labels assigned to the topics showed that some topics are more dominant in the corpus. In summary, observations are as follows: topic labeling can be considered in combination with topic trends over time. In this way, the correlation between topic trends may help in both label evaluation and label assignment. Comparing the inferred topics on monthly and weekly data shows that the earlier one is more trustworthy (less correlation between trends over time). Capturing all noised in particular topics may lead us to topic based noise removal methods in a corpus.

Acknowledgments

Special thanks to prof. An for the excellent comments.

References

1. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1999) 50–57
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3** (2003) 993–1022
3. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century american newspaper. Journal of the American Society for Information Science and Technology **57**(6) (2006) 753–767
4. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. (2009) 288–296
5. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the

North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 100–108
6. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of latent semantic analysis **427**(7) (2007) 424–440
7. Wei, Y., Taylor, A., Yossinger, N.S., Swingewood, E., Cronbaugh, C., Quinn, D.R., Singh, L., Martin, S.F., Berkowitz, S., Collmann, J., et al.: Using large-scale open source data to identify potential forced migration. In: Workshop on Data Science for Social Good at KDD. (2014)