TED Talks' Topic Variation Utilizing a Dynamic Topic Modeling Approach

Seung-Ho Ryu[a] • Young Eun Jeon[a] • Jung-In Seo[a]

[a] Department of Data Science, Andong National University, Andong, Korea

1. Abstract

Motivation

- TED is a platform that shares innovative ideas across various fields such as technology, medicine, and design to inspire and provoke change in people's thoughts and behaviors.
- By capturing the changes of topics in TED Talks, we can not only understand global trends and public interests but also identify a current social issue or situation.

Purpose

- The goal of this study is to identify the topic changes over time by applying the dynamic topic modeling (DTM) to a TED Talks dataset (Source : https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset).

Conclusion

- Content planners and marketing strategists can leverage our analysis results to predict future trends and select topics that are likely to captivate the audience.

2. Application

- Our framework for analysis on a TED Talks dataset is shown in Figure 1.

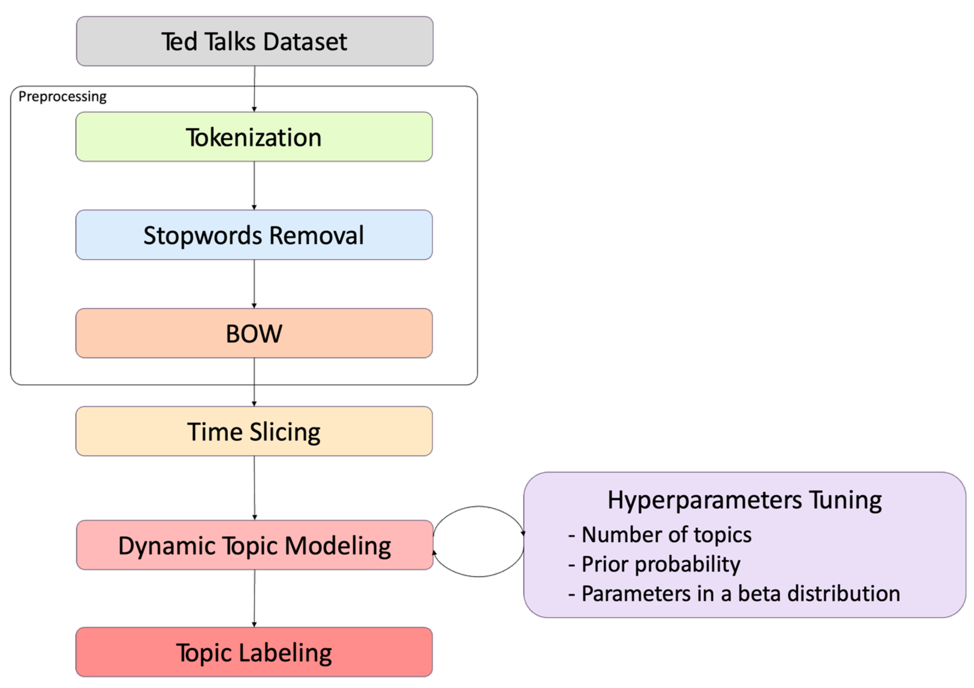

Figure 1. Framework for analysis on the TED Talks dataset

2-2 TED Talks

- Among talks available on TED.com from January 1, 2017 to December 31, 2019, the transcripts for talks translated into English are employed for analysis.

2-3 Preprocessing

- **Tokenization**
  - Tokenization is the process of splitting the text into small units called tokens.
  - Tokens can be individual words, phrases, or other meaningful elements.
- **Stopwords removal**
  - Stopwords removal is the process of eliminating high-frequency but unimportant tokens such as morphemes, prepositions, etc.
- **Bag of Words (BOW)**
  - BOW is an embedding method that represents text data as numerical vectors based on the frequency of each word.

2-4 DTM

I. For each topic $k \in \{1, \dots, K\}$:

a. Draw $\boldsymbol{\eta}_{t,k} = (\eta_{t,k,1}, \dots, \eta_{t,k,V_t}) | \boldsymbol{\eta}_{t-1,k} \sim N_k(\boldsymbol{\eta}_{t-1,k}, \sigma^2 I)$

II. Draw $\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1} \sim N_k(\boldsymbol{\alpha}_{t-1}, \delta^2 I)$

III. For each document $d \in \{1, \dots, D\}$:

a. Draw $\boldsymbol{\tau}_d = (\tau_{d,1}, \dots, \tau_{d,K}) \sim N_k(\boldsymbol{\alpha}_t, a^2 I)$

b. For each word $n \in \{1, \dots, N_d\}$:

$$\text{Compute } \boldsymbol{\theta}_{t,d} = \frac{exp(\tau_d)}{\sum_{i=1}^{K} exp(\tau_{d,i})} \Longrightarrow \text{Draw a topic assignment } Z_{t,d,n} \sim Multinomial(\boldsymbol{\theta}_{t,d})$$

$$\text{Compute } \boldsymbol{\beta}_{t,z_{t,d,n}} = \frac{exp(\eta_{t,z_{t,d,n}})}{\sum_{i=1}^{V_t} exp(\eta_{t,z_{t,d,n},i})} \Longrightarrow \text{Draw a word } W_{t,d,n} \sim Multinomial(\boldsymbol{\beta}_{t,z_{t,d,n}})$$

- ✓ t : Time slice
- ✓ K : Number of topics
- ✓ D : Number of documents
- ✓ N_d : Number of words in document d
- ✓ V_t : Number of unique words in the entire corpus at time t
- ✓ $\boldsymbol{\theta}_{t,d}$: Topic proportion vector in document d at time t
- ✓ $\boldsymbol{\beta}_{t,k}$: Probability vector of the word appearance of the topic k at time t
- ✓ $\boldsymbol{\tau}_d$: Latent variable used to calculate $\boldsymbol{\theta}_{t,d}$
- ✓ $\boldsymbol{\eta}_{t,k}$: Latent variable used to calculate $\boldsymbol{\beta}_{t,k}$
- ✓ $\boldsymbol{\alpha}_t$: Parameters of document-topic distribution at time t

3. Results

3-1 Composition of topics over time

- Figure 2 visualizes a topic prevalence with the word composition of a topic for each year.
  - ✓ A circle size represents the topic's relative prevalence in the entire corpus.
  - ✓ A distance between the circles indicates a similarity of topics.
- Figure 3 shows the top 10 words with the highest probability of occurrence.
  - ✓ A length of bar represents a word occurrence probability in a topic.

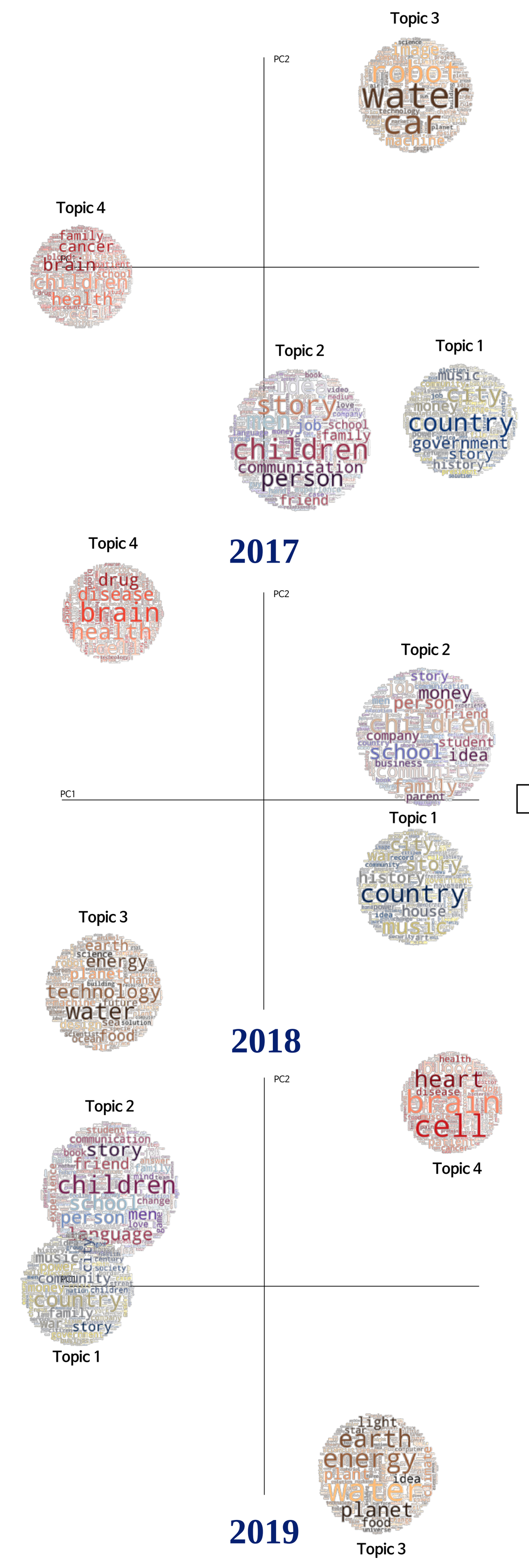

Figure 2. Overall compositions of topics for each year

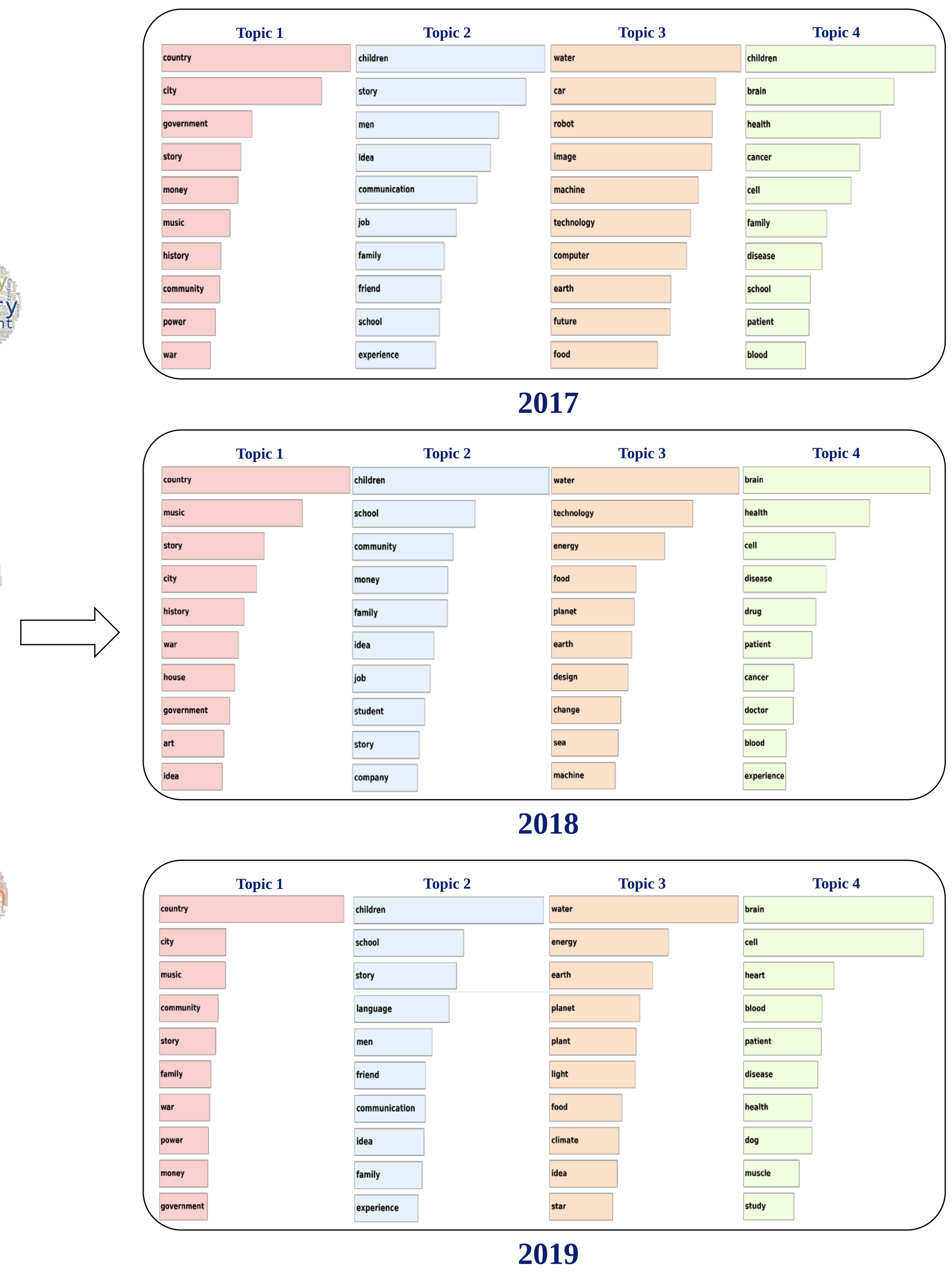

Figure 3. Top 10 words for each year

- For all years, the circle size of Topic 2 is the largest, which implies that the importance of this topic is the highest.
- For all years, considering the relationship between topics based on the distance, the similarity between Topic 2 and Topic 3 is the lowest.
- The word "*government*" in Topic 1 shows a decreasing probability of occurrence over time.

3-2 Topic Similarity

- Figure 4 shows similarities between different topics in all years.

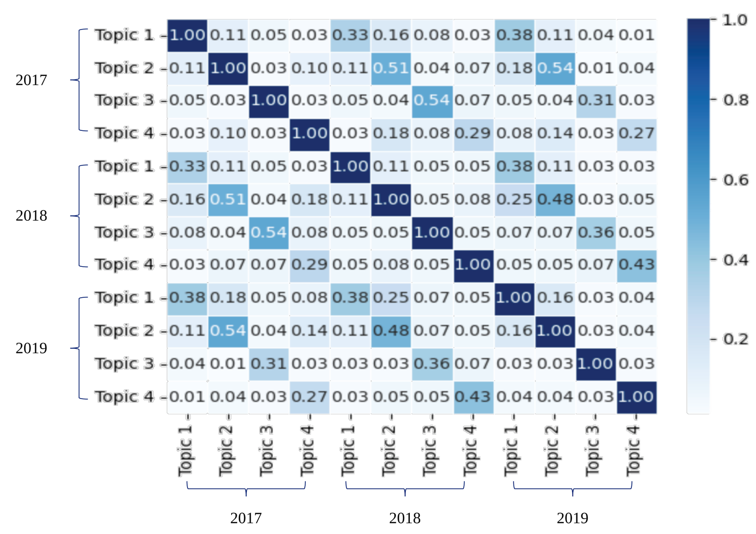

Figure 4. Heatmap of topic similarity using Jaccard distance

- Similarity is represented as a value between 0 and 1.
  - ✓ A value close to 0 indicates low similarity between two topics.
  - ✓ A value close to 1 indicates high similarity between two topics.
- Figure 4 shows that most topics exhibit low similarity, which reveals that the DTM effectively separates each topic and the topics reflect various aspects of a TED Talks dataset.

3-3 Topic Labeling

- Based on Figures 2-4, the topics can be labeled as follows:

	2017	2018	2019
Topic 1	Cultural Narratives and history	Cultural and political change	Community, governance, and social change
Topic 2	Communication and relationships among children	Educational interactions and socioeconomic factors in childhood	Social dynamic and language development in childhood
Topic 3	Future technologies and environmental sustainability	Technology innovation and environmental solutions	Exploring planetary systems and environmental impact
Topic 4	Advances in infant brain and cancer research	Neurological research and scientific enquiry of disease	Advances in cardiac and genetic research

4. Conclusions

- **Content Strategy Planning**: Identifying popular topics and those that receive less interest is beneficial for determining future lecture topics.
- **Future Prediction and Planning**: Based on the trends derived from dynamic topic modeling, it is possible to predict future changes in topics. This is beneficial for planning the direction of TED Talks.
- Such analysis can enrich TED Talks content and greatly contribute to meeting the diverse needs of the audience.