Graph-based Trajectory Visualization for Text Mining of COVID-19 Biomedical Literature

Yeseul Jeon[1,2], Dongjun Chung[3], Jina Park[1,2] and Ick Hoon Jin[1,2]

Abstract

Since the emergence of the worldwide pandemic of COVID-19, relevant research has been published at a dazzling pace, which makes it hard to follow the research in this area without dedicated efforts. It is practically impossible to implement this task manually due to the high volume of the relevant literature. Text mining has been considered to be a powerful approach to address this challenge, especially the topic modeling, a well-known unsupervised method that aims to reveal latent topics from the literature. However, in spite of its potential utility, the results generated from this approach are often investigated manually. Hence, its application to the COVID-19 literature is not straightforward and expert knowledge is needed to make meaningful interpretations. In order to address these challenges, we propose a novel analytical framework for effective visualization and mining of topic modeling results. Here we assumed that topics constituting a paper can be positioned on an interaction map, which belongs to a high-dimensional Euclidean space. Based on this assumption, after summarizing topics with their topic-word distributions using the biterm topic model, we mapped these latent topics on networks to visualize relationships among the topics. Moreover, in the proposed approach, the change of relationships among topics can be traced using a trajectory plot generated with different levels of word richness. These results together provide a deeply mined and intuitive representation of relationships among topics related to a specific research area. The application of this proposed framework to the PubMed literature shows that our approach facilitates understanding of the topics constituting the COVID-19 knowledge.

Introduction

As of June 2021, more than 172 million cases have occurred, with more than 3.7 million deaths, across the world (Dong et al. 2020). In spite of emergence of COVID-19 vaccines, most countries still suffer from COVID-19 and we are still far from overcoming it. Tremendous efforts to understand and overcome this fatal virus have led to the rich amount of biomedical literature about COVID-19. As of June 2021, searching with the term "COVID-19" in the PubMed biomedical literature database (https://pubmed.ncbi.nlm.nih.gov) resulted in more than 140,000 publications. These literature covers a wide range of topics, including general description, mechanism, transmission, diagnosis, treatment, prevention, case report, and forecasting, among others, as profiled and categorized in LitCovid, a biomedical literature database dedicated to COVID-19 (Chen et al. 2020). In order to facilitate understanding of relevant mechanisms and come up with effective treatment and preventive strategies, it is critical to follow up and digest these publications that are being published in a fast phase. However, given the rich volume of COVID-19 literature and its rapid publication phase, it is practically impossible for biomedical experts to trace all of these literature manually in real time. Statistical and computational approaches, especially text mining, can be a powerful solution for the researchers to address this challenge. Since literature is unstructured text data,

a document is usually first summarized with the corpus, representing word counts. The *tf-idf* approach is most commonly used to summarize the document with the words' frequency (Chowdhury 2010). In spite of its popularly, this approach is still naive and limited in the sense that each single word is treated independently and relationships among them are totally ignored. However, such latent relationships can provide invaluable information that leads to effective modeling of literature and ultimately deeper understanding of the literature. Therefore, it is of great interest to cluster words effectively and understand meaning of these clusters.

Topic Modeling In the text mining area, it is of great interest to effectively cluster words or documents and topic modeling is the most well-known approach for this purpose. It essentially extracts topics based on the words' distribution in each document and various algorithms have been proposed for topic modeling. The first-of-its-kind model can be found

[1]Department of Statistics and Data Science, Yonsei University, Seoul, South Korea
[2]Department of Applied Statistics, Yonsei University, Seoul, South Korea
[3]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio 43210, United States of America

Corresponding author:
Ick Hoon Jin, Department of Applied Statistics, Yonsei University, Seoul, South Korea
Email: ijin@yonsei.ac.kr

in Deerwester et al. (1990), which used singular value decomposition with a word-by-document matrix to construct the semantic structure. Hofmann (2001) improved it by using a generative latent class model that essentially replaces word counts with word occurrence probabilities. Blei et al. (2003) made an important step in this direction and proposed a generative probabilistic model for word set, called the Latent Dirichlet Allocation (LDA). LDA assumes each document is structured in the context of topics while words are allocated to these topics. The topics are regarded as latent variables and characterized by words with probabilities assigned to each topic. To estimate latent topics, LDA uses Bayesian statistical methods such as Expectation-Maximization (EM) algorithm and variational inference to construct the posterior distributions of output: (1) topic distributions and (2) topic-word distributions. First, the topic distribution models how likely each topic is relevant to a given document. Second, the topic-word distribution represents the probability how likely each given word is associated with a particular topic. A topic model has been widely used and applied in diverse text data (Jiang et al. 2015). For example, Amado et al. (2018) applied the LDA topic modeling to articles published in Science Direct's database as an attempt to identify Big Data trends in marketing. This approach revealed an increasing interest in Big Data pertaining to marketing, grouping three terms related to Big Data and eight topics associated with marketing. On the other hand, there have been attempts to apply topic modeling to the literature of biological or biomedical text data. Liu et al. (2016) applied topic modeling to biological literature and showed that topic modeling can be a practical solution for the bioinformatics research as well.

Since COVID-19 is a hot topic and its research is active and ongoing, some articles are released without full context but with only abstract. Abstracts of the COVID-19 literature are categorized as short text. Unfortunately, with short texts, conventional topic models including LDA often suffer from poor performance due to the sparsity problem, which causes inferior topic inferences. There have been some attempts to handle this problem by using additional assumptions for short text data. For instance, Zhao et al. (2011) and Lakkaraju et al. (2012) trained a topic model on tweet data as a mixture of unigrams (Nigam et al. 2000), where a document is considered as a word set drawn independently from a single topic. Likewise, Gruber et al. (2007) assumed that the words within the same sentence share the same topic. Although these constraints may help alleviate the data sparsity problem caused by short text data, this improvement was achieved by using more information to capture multiple topics from each document. Moreover, these assumptions tend to result in peaked posteriors of topics in a document, making the model susceptible to overfitting (Blei et al. 2003). Yan et al. (2013) made important progress for modeling of short text data, so-called Biterm Topic Model (BTM). Unlike the LDA, BTM replaced words with bi-terms, where a bi-term is defined as a set of two words occurring in the same document. This approach attempted to make up for the lack of words in a short text by pairing two words, thereby creating more words in the document. This allows considering that if two words are mentioned together, they are more likely to belong to the same topic. Some studies apply BTM to short text data such as micro-blog (Li et al. 2016), large-scale short texts, and tweets for clustering and classification (Chen and Kao 2017).

Topic Model Visualization Generally text data analyses end up with listing topics and associated words and it requires expert knowledge to make interpretation of these results. Hence, visualization of topic modeling results can be a powerful solution as it can reduce the interpretation burden and there have been several attempt to visualize the output of topic models. For instance, Andrews et al. (2002) clustered the document by using a modified weighted Voronoi diagram that represents the number of documents. Cao et al. (2010) visualized multifaceted relations of documents by deriving a density map that organized the literal meaning of words of health care in advance. Lafferty and Blei (2006) visualized the topic relationship as a static graph. Newman et al. (2010) revealed the semantic structure of document collection and visualized the relationship between topic and document. Gretarsson et al. (2012) provided a web-based system that could interactively visualize topics, allowing for better understanding of documents. Essentially, these approaches consider documents and topics as node-edge relationships on a network and estimate topic relationships based on similar documents. Another approach to visualize the topic model utilizes the information about topics and their associated words. For example, Sievert and Shirley (2014) proposed LDAvis, a web-based interactive tool for visualizing topics estimated using LDA. This framework quantify the relevance between word and topic, and it calculate the probability of a given word being affiliated with a given topic. In this approach, topic-word relationships can be investigated by checking affiliated words for the selected topic.

In spite of these previous attempts to visualize topic modeling results, it is important to note that in the real world, it is hard to find the data that contain independent topics towards one subject but instead multiple topics are often inter-correlated. Therefore, it might not be sufficient to show each topic separately without considering relationships among topics. Moreover, visualization can be significantly improved when the relationships between topics and the relationships between words are modeled simultaneously. Specifically, this joint modeling of multi-layered relationships (topic-topic, word-word, and topic-word) can provide global and local representation of topics at the same time. In this context, Mei et al. (2008) tried to identify topical communities by combining a topic model with a social network. Specifically, it estimated topic relationships given the known network structure and tried to separate out topics based on connectivity among them. However, in this approach, the network information needs to be given to construct and visualize the topic network, which is not a trivial task in practice. Our study aims to overcome these limitations and estimate the topics' relationship and visualize their dependency structure simultaneously, by considering interaction among topics using network models. In addition, we also aim to overcome a static topic network by dynamically representing fundamental associations among topics through tracking the change in relationships.

Our Analytic Framework Our analytic framework consists of the following four steps. In the first step,
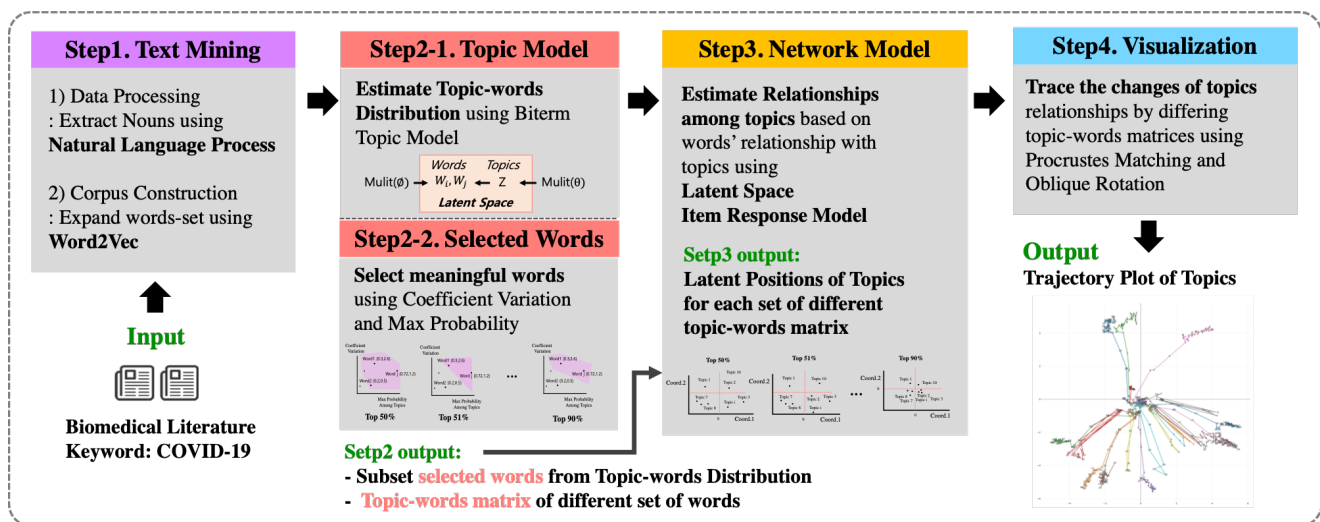
Figure 1. Illustration of the proposed approach for effective visualization of topic modeling results. Step 1: Processing of text mining results with natural language processing and construction of the word corpus. Step 2: Extraction of topics, each of which is defined by its own word distribution estimated using the Biterm Topic Model (BTM). Step 3: Visualization of the topics and their word distributions as networks, using the latent space item response model (Jeon et al. 2021). Step 4: Visualization of relationships among topics using the trajectory plot showing transition of topics' coordinates.

text mining is conducted to extract nouns and corpus is constructed for modeling literature data. In the second step, we use BTM (Yan et al. 2013) to estimate topics and their associated words. In the third step, latent topic positions are estimated based on the distance that measures similarity between a topic and a word. In the last step, we visualize their relationships by tracking the transition of latent positions of topics. This approach represents relationships among topics on the latent positions and this facilitates understanding of distinct properties of the topics. Specifically, the more topic A shares words with topic B, the more likely topic A is located closer to topic B, which nicely depicts the topic relationships identified by BTM.

This article is composed of three parts. First, we will briefly summarize the contributions of this paper and then introduce our methods. Second, we will describe the topic modeling, specifically the BTM, an alternative topic modeling to LDA for the purpose of tackling the short text problem. Third, we will introduce the latent space item response model (Jeon et al. 2021), which estimates associations between items based on the latent distance between item and response. Finally, we will apply our approach to the COVID-19 literature to evaluate and demonstrate the usefulness of our approach.

Our Contributions

There are three key contributions of this work. First of all, our paper visualizes the result of the topic models by mapping unobserved interactions among topics onto the interaction map based on the latent distance between topics and words, where the interactions among topics are quantified by their interactions with words. This embedded derivation of topic relationships will reduce the burden of data analysts because it does not require prior knowledge about relationships between topics, e.g., a connectivity matrix. Instead, our approach utilizes the topic-word matrix for calculating the

latent positions of each topic, where each cell corresponds to the probability of a word's priority linkage to a topic. Since our approach uses continuous values ranging from 0 to 1 for the modeling purpose, it can quantify closeness between topics by reflecting degrees of linking of words to topics and, hence, using latent positions for topics, we could represent relationships among topics more precisely.

Secondly, we visualize the topic relationship as a trajectory plot to detect the change of interactions of topics over different set of words. This feature has two important properties; (1) it could measure the main location of the topic, which is steadily positioned in a similar place in spite of differing the network structure; and (2) we could infer popular topics mentioned across articles and recently emerging topics by scanning the latent position of each topic. Specifically, if a particular topic shares most of the words associated with it with other topics, it is more likely to be located in the center on the latent interaction map. In contrast, if a specific topic consists mostly of words unique to that topic (e.g., a rare topic or an independent topic containing its own referring words), it is more likely to be located away from the center on the interaction map. For example, in the context of COVID-19, it is more likely that common subjects like 'outbreak' and 'diagnosis' are located in the center while more specific subjects like 'Cytokine Storm' are located more outside on the interaction map.

Finally, this approach helps organize an tremendous amount of literature and mine underlying relationships among topics based on the literature. The topic network that visualizes the relationship of topics gives researchers more intuition to set out their studies. For instance, if some researchers want to investigate a specific topic, say Topic A, our framework can assist them by providing information answering the three following questions: (1) Which set of words are associated with Topic A? Researchers can obtain this information from the BTM results. In addition, since we extract meaningful words that distinctively represent each

Step 2-1. Biterm Topic Modeling Outline

Step 2-2. Extract sets of meaningful words
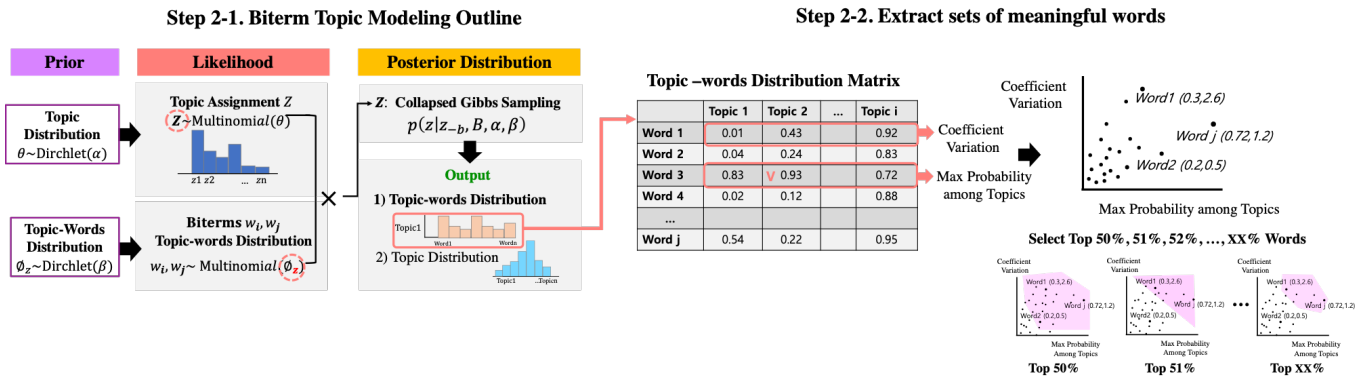
Figure 2. Detailed description of Step 2 of the proposed approach. **Figure of the left: Step 2-1.** The likelihood of BTM is obtained as the joint distribution of topic-word relationships and topics. BTM provides the topic-word distribution $\phi_{w|z}$ and the topic distribution θ_z. **Figure on the right: Step 2-2.** Through the BTM, we can construct the X matrix, of which dimension is $N \times P$, where each cell is composed of probability for the degree of association of each word with each topic. We extract sets of meaningful words representing characteristics of topics using two statistical measurements, including coefficient variation and maximum probability from each row of matrix X.

topic's meaning, our approach could further support this investigation with a more refined word set. (2) Are there any other topics related to Topic A for extending and elaborating research? Researchers can answer this question by intuitive interrogating the final visualization. (3) Is Topic A is a common or specific topic? Since we could trace change of latent topic positions as a function of words set, our method provides relevant insights through topic locations on the interaction map. Specifically, researchers can consider Topic A to be common if it is located in the center, and to be specific if it is located away from the center.

Methodology

Graph-based Visualization and Clustering for Text Mining

We developed a novel framework for effective visualization of topic modeling results, which positions topics on an Euclidean latent space. This framework consists of the following four steps, as illustrated in Figure 1. In the first step, we implement text mining with natural language processing and construct the word corpus by expanding a set of words using the `word2vec` model (Griffiths and Steyvers 2004). In the second step, based on this word corpus, we extract the topic-word distribution, where each topic is characterized with the corresponding topic-specific distribution of words estimated using the BTM. Using this topic-word distribution, we can extract meaningful words that affect topic characteristics. In the third step, the topic-word relationship is estimated using the latent space item response model (Jeon et al. 2021), which provides latent topic positions. Finally, in the fourth step, we can visualize relationships among topics with topic-specific traces.

Step 1: Processing Text Data with Text Mining
The text data from biomedical papers will be input for this step. We extract information from the text using a morphology analysis, one type of natural language processing techniques. Specifically, it splits each word with suffix to identify the base element unit of a term. Among

the basic units of words, we first extract nouns in their most basic forms. This set of nouns is called a corpus. We further expand the corpus by adding relevant words using `word2vec`, which vectorizes distances among words in the Euclidean latent space according to their similarity in meaning. We extract neighboring words based on these distances and enlarge the word set by collecting these words.

Step 2-1: Estimating Latent Topics using Topic Modeling Following Step 1, it is necessary to understand the overall semantic structures of a given text. This step summarizes the latent structures, called topics, by identifying topics and estimating their corresponding clusters of words. That is, we can estimate topics and their distributions of words using topic modeling.

Biterm Topic Modeling for Short Text The abstract is an excellent source to understand the overall text. Since most abstracts are limited to 200 words, they could be regarded as short texts. Therefore, we use the BTM for literature mining. To implement BTM, we first need a word corpus, which is the output of the previous step. Then, we extract biterms from each paper, which is input for the BTM.

Likelihood of Biterm Topic Model The BTM is based on the following assumption describing how bi-terms and topics are jointly related:

- Two words are assumed to belong to some hidden clustered set of words.

- It is assumed that there are hidden topics, each of which contains similar meaning of words.

- It is assumed that those hidden clustered set of words are hidden topics.

Based on these assumptions, as shown in Figure 2, the likelihood of BTM consists of both topic-word distribution and topic distribution. Therefore, we need two sets of parameters to estimate the topic distribution θ and the topic-word distribution ϕ_z. The whole likelihood of BTM is constructed as follows. First, the prior distribution for words
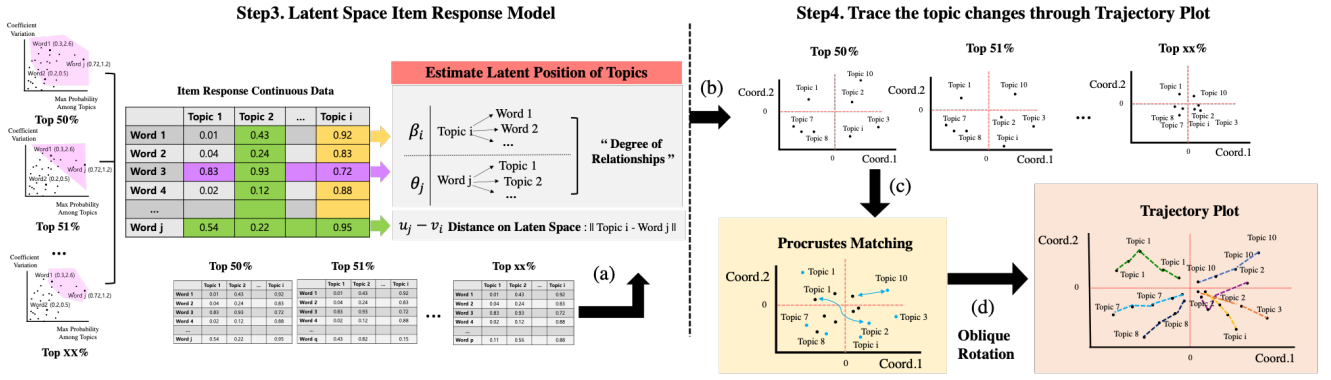
Figure 3. Detailed description of Steps 3 and 4 of the proposed approach. Figure on the left: Step 3. (a) By fitting the LSIRM models with various sets of matrices ($\mathbf{X_i}$) as input, we can obtain matrices of A_i composed of coordinates of each topic. **Figure on the right: Step 4. (b)** topics' latent positions are visualized for various input matrices $\mathbf{X_i}$) obtained from Step 2-2. **(c)** Topics for different input matrices are aligned to the same quadrant using the Procrustes matching. **(d)** The axis are improved using the oblique rotation and transition of the topic's latent positions as a function of the number of words are visualized as a trajectory.

(ϕ_z) is set to Dirichlet distribution with hyper-parameter β while the prior distribution for topics is set to Dirichlet distribution with hyper-parameter α. Next, we represent the topic with the latent variable z, which follows Multinomial distribution with parameter θ. Likewise, each word follows Multinomial distribution with paramereter ϕ_z so that each word can be generated from a specific topic. Therefore, there are three parameters to estimate, including ϕ_z, θ, and z. The likelihood construction process can be summarized as follows:

Step 1. Draw a whole topic distribution from $\theta \sim$ Dirichlet(α).

Step 2. For each biterm b from biterm set B, calculate to which among latent topics z those two words belong: $z \sim$ Multinomial(θ)

Step 3. Draw a topic-word distribution of topic z from $\phi_z \sim$ Dirichlet(β).

Step 4. Draw probabilities for two words from the topic-word distribution corresponding to the selected topic: $w_i, w_j \sim$ Multinomial(ϕ_z)

After the above steps, the joint likelihood of all biterm set B is calculated as:

$$p(B) = \prod_{i,j} \sum_z \theta_z \phi_{i|z} \phi_{j|z} \quad (1)$$

Collapsed Gibbs Sampling Algorithm The conditional posterior for a latent topic z, $p(z|z_{-b}, B, \alpha, \beta)$, is given as:

$$p(z|z_{-b}, B, \alpha, \beta) \propto (n_z + \alpha) \frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2} \quad (2)$$

where n_z is the number of times that the biterm b is assigned to topic z, z_{-b} refers to the topic without biterm b, and $n_{w|z}$ is the number of times of the word w assigned to the topic z.

Because a direct application of the Gibbs sampling for $p(z|z_{-b}, B, \alpha, \beta)$ sometimes can result in lack of convergence due to dependency between variables, we use

the collapsed Gibbs sampling to constrain unnecessary parameters by integrating out (Liu 1994). In particular, because the prior distributions are Dirichlet distributions, ϕ_z and θ can be integrated out. After some iterations, we can construct distribution of $\phi_{w|z}$ and θ_z with estimated statistics $n_{w|z}$ and n_z, given as follows:

$$\phi_{w|z} = \frac{n_{w|z}}{\sum_w n_{w|z} + M\beta}, \quad \text{and} \quad \theta_z = \frac{n + \alpha}{|B| + K\alpha} \quad (3)$$

Algorithm 1 Gibbs sampler for BTM

Input: the number of topics; K, hyper-parameters; α, β, biterm set; B
Output: distribution ; $\phi_{w|z}$ and θ_z

1: initialize topic assignments randomly for all the biterms
2: **for** iteration= $1, 2, \ldots, N$ **do**
3:     **for** $b = 1, 2, \ldots, B$ **do**
4:         sample z_b from $p(z|z_{-b}, B, \alpha, \beta)$
5:         update statistics $n_{w|z}$ and n_z
6:         compute the parameter $\phi_{w|z}$ and θ_z
7:     **end for**
8: **end for**

Step 2-2: Selecting Key Words from Topics

After Step 2-1, we can obtain the topic-word distribution $\phi_{w|z}$ and the topic distribution θ_z. Each topic contains words and their corresponding probabilities so it is meaningful to compare each topic based on their word distributions. The simplest way to distinguish topics is to compare word memberships between topics. Since the output of each topic-word distribution includes all words, we might not be able to determine characteristics of topics if we use all the words that topics share. Therefore, we select only meaningful words from each topic to estimate the relationships among topics based on their representative words.

According to Figure 3, through the BTM, we can construct the $\mathbf{X_i}$ matrix with dimension $N \times P$, where N denotes the number of words, P refers the number of topics, and each cell represents probabilities to belong to each topic. To determine a criterion to choose words, we calculate

coefficient variation and maximum probability from each row of matrix $\mathbf{X_i}$. Each value represents the variation of word probabilities among different topics and the degree of word's linkage to a particular topic. Since BTM results in topic-word distribution where each word's probability is measured within each topic, we need to adjust their variation rather than simply calculating a variance. Therefore, we divide the standard deviation by the mean of word's probabilities from each topic.

There are two reasons for choosing a coefficient variation and maximum probability from each row of matrix $\mathbf{X_i}$ as a criterion to select words. First, it is expected that important words have large variation in probabilities among topics because if a word has low variation across topics, it is likely that word does not represent any topic specifically. Therefore, meaningful words can be selected based on their variations and using coefficient variation can further scale the words' dispersion among topics. Second, in order to be a meaningful word, it should also have high probability in at least one topic. For example, if a word has high variation but only low possibilities across topics, this word still cannot differentiate topics. Therefore, we can effectively characterize the topic by selecting a word with high probability in at least one topics and those with large variation.

Step 3: Estimating Topic Relationships Using Latent Space Item Response Model

In Step 3, we estimate interactions among topics and visualize their relationships by mapping on the interaction map. Hoff et al. (2002) proposed the latent space model, which expresses a relationship between actors of a network in an unobserved "social space", so-called latent space. Inspired by Hoff et al. (2002), Jeon et al. (2021) proposed the latent space item response model (LSIRM) that viewed item response as a bipartite network and estimated the interaction between respondents and items using the latent space. LSIRM models two parts: an attribute part and an interaction part. In the attribute part, LSIRM estimates how much respondents respond to certain items and how many items are answered by some respondents. In the interaction part, LSIRM estimates the interaction between items and respondents, along with the latent position of each item and respondent on the interaction map.

Given our goal to estimate the relationship among topics and visualize their latent positions on the interaction map based on their associated words, we use LSIRM with $\mathbf{X_i}$ from Step 2 as a bipartite network, where an item indicates a topic and respondents refers to words. However, the original LSIRM proposed by Jeon et al. (2021) cannot be directly applicable here because it was designed for binary item response dataset, where each cell in the item response data is binary valued (0 or 1). On the contrary, here our input data $\mathbf{X_i}$ has continuous probabilities indicating how likely each word belongs to each topic. Therefore, in order to apply LSIRM to our input data $\mathbf{X_i}$, we need to expand the Jeon et al. (2021) model to a Gaussian version, which is described in detail below.

The modified Gaussian version of LSIRM can be written as

$$
\begin{aligned}
x_{j,i} \mid \mathbf{\Theta} &= \theta_j + \beta_i - ||\mathbf{u}_j - \mathbf{v}_i|| + \epsilon_{j,i}, \\
\epsilon_{j,i} &\sim \mathrm{N}(0, \sigma^2).
\end{aligned}
\tag{4}
$$

where $x_{j,i}$ indicates the probability of word j belongs to topic i, for $i = 1, \cdots, P$ and $j = 1, \cdots, N$. Because the original LSIRM use logit link function to handle the binary data, here we use the linearity assumption between $x_{j,i}$ and the attribute part with the interaction part. We add an error term $\epsilon_{j,i} \sim N(0, \sigma^2)$ to satisfy the normality equation. We use the notation $\mathbf{\Theta} = \{\boldsymbol{\theta} = \{\theta_j\}, \boldsymbol{\beta} = \{\beta_i\}, \mathbf{U} = \{\mathbf{u_j}\}, \mathbf{V} = \{\mathbf{v_i}\}\}$, and $||\mathbf{u}_j - \mathbf{v}_i||$ represents the Euclidean distance between latent positions of word i and topic j. Here, the shorter distance between $\mathbf{u_j}$ and $\mathbf{v_i}$ implies the higher probability that word j links to in topic i. Therefore, latent positions of topics can be estimated based on the distances with words.

Given the model described above, we use Bayesian inference to estimate parameters in the Gaussian version of LSIRM. We specify prior distributions for the parameters as follows:

$$
\begin{aligned}
\beta_i | \tau_\beta^2 &\sim \mathrm{N}(0, \tau_\beta^2), \quad \tau_\beta^2 > 0 \\
\theta_j | \sigma^2 &\sim \mathrm{N}(0, \sigma_\theta^2), \quad \sigma^2 > 0 \\
\sigma^2 &\sim \mathrm{Inv\text{-}Gamma}(a, b), \quad a_> 0, \quad b > 0 \\
\sigma_\theta^2 &\sim \mathrm{Inv\text{-}Gamma}(a_\sigma, b_\sigma), \quad a_\sigma > 0, \quad b_\sigma > 0 \\
\mathbf{u_j} &\sim \mathrm{MVN}_d(\mathbf{0}, \mathbf{I}_d) \\
\mathbf{v_i} &\sim \mathrm{MVN}_d(\mathbf{0}, \mathbf{I}_d).
\end{aligned}
$$

where $\mathbf{0}$ is a d-vector of zeros and $\mathbf{I}_d$ is the $d \times d$ identify matrix. We fixed τ_β^2 as constant value.

The posterior distribution of LSIRM is

$$
\begin{aligned}
\pi(\mathbf{\Theta}, \sigma^2 | \mathbf{X}) \propto &\prod_j \prod_i \mathbb{P}(x_{ji} | \mathbf{\Theta}) \\
&\times \prod_j \pi(\theta_j | \sigma_\theta^2) \pi(\sigma_\theta^2) \prod_i \pi(\beta_i) \\
&\times \prod_j \pi(\mathbf{u_j}) \prod_\mathbf{i} \pi(\mathbf{v_i}) \pi(\sigma^2)
\end{aligned}
$$

and we use Markov Chain Montel Carlo (MCMC) to estimate the parameters of LSIRM. In this way, we can obtain latent positions of $\mathbf{u_j}$ and $\mathbf{v_i}$ on the interaction map R^d. Since we are interested in constructing the topic network, we utilize $\mathbf{u_j}$ and make it as matrices of $\mathbf{A}_i$.

Step 4: Visualization Using A Trajectory Plot

After proceeding with the LSIRM model with various sets of matrices $\mathbf{X_i}$, we could obtain matrices of $\mathbf{A}_i$, composed of coordinates of each topic. In order to further improve interpretation of relationships among topics, we trace how their latent positions change as a function of word sets. Specifically, we compare topics' latent positions $\mathbf{A}_i$ from each matrix $\mathbf{X_i}$ with the following steps.

First, we implement a Procrustes matching two times: (1) within the MCMC samples generated from LSIRM, to tackle the invariance property (so-called within-matrix matching); and (2) for the estimated matrices, to locate topics on the same quadrant (so-called between-matrix

matching). Second, we take average of the distances of topics' latent positions from the origin, to measure the degree of dependency structure. Specifically, the longer distance of latent positions implies the stronger dependency of the network. To locate matrix $\mathbf{X}_i$ onto the same quadrant, we choose the baseline matrix, which maximizes the dependency structure among topics. It helps nicely show the change of topics' latent positions because those rotated positions $\mathbf{A}_i$ from each matrix $\mathbf{X}_i$ are based on the most stretched out network from the origin. Note that the baseline matrix is $\mathbf{A}_{\max}$. Finally, we rotate the axes to improve the interpretability of the relationship among topics using oblique factor rotation (Jennrich 2002). After implementing these steps (which correspond to the steps **(c)** and **(d)** in Figure 3), we visualize relationships among topics by tracing the coordinates of each topic.

Procrustes Matching As fitting the LSIRM model, we collect MCMC samples for each topic's latent positions. Note that there exists multiple possible realizations for latent positions because the distances between pairs of respondents and items, which are included in a likelihood function, are invariant under rotation, translation, and reflection (Hoff et al. 2002; Shortreed et al. 2006). In order to tackle this invariance property for determining latent positions, we implement within-matrix Procrustes matching (Borg and Groenen 2005) as post-processing of MCMC samples.

After implementing the within-matrix matching for the matrix $\mathbf{A_i}$, we additionally execute the between-matrix Procrustes matching based on the baseline matrix $\mathbf{A}_{\max}$ to compare the transition of the topic's latent positions for the different input matrices $\mathbf{A_i}$. We assign $\mathbf{A}_{\max}$ to the matrix that maximizes $\mathbf{A_i}$ dependency. Finally, we can obtain the re-positioned matrices $\mathbf{A_i^*}$, which still maintain the dependency structure among topics but located in the same quadrant.

Oblique Rotation With the oblique rotation, the interpretability of axes can be further improved and topics can be categorized based on these axes. For this purpose, we apply the `oblim` rotation (Jennrich 2002) to the estimated topic position matrix $\mathbf{A_i^*}$, using the R package `GPAroation` (Bernaards and Jennrich 2005). We denote the rotated topic position metric by $\mathbf{B_i}$. To interpret the trajectory plot showing traces of topics' latent positions, we extract the rotation information matrix ($\mathbf{R}$) resulting from an oblique rotation as the baseline matrix $\mathbf{B_{base}}$. Then, we multiply each matrix ($\mathbf{B_i}$) by the rotation matrix ($\mathbf{R}$) to plot the topics' latent positions.

Real Data Application

COVID-19 Biomedical Literature We applied our workflow to the COVID-19 literature to investigate the latent semantic structure of COVID-19 literature. In our implementation, we first downloaded the COVID-19 articles published between December 1st, 2019 and August 3rd, 2020 from the PubMed database (https://pubmed.ncbi.nlm.nih.gov) . Specifically, we collected articles of which titles contain "coronavirus2", "covid-19" or "SARS-CoV-2" and this resulted in total 35,585 articles. After eliminating articles without abstracts (i.e., only titles or abstract keywords), our final text data contained the total 15,015 documents.

Corpus To construct the corpus, we used abstract keywords that concisely capture messages delivered by the paper. To achieve the richer corpus, we also used the `word2vec` (Mikolov et al. 2013a) to train against relationships between nouns from the abstract and the abstract keywords. Specifically, `word2vec` extracted nouns from abstracts, which were embedded near to the abstract keywords, and added those selected words to the corpus. To train the words' network, Mikolov et al. (2013b) suggested the negative-sampling approach, which fits the network of words by training the near words and unrelated words, and this approach was reported to be efficient in vectorizing the words' relationships. Goldberg and Levy (2014) expanded the negative-sampling strategy to model the words and the contexts through the joint modeling of words and contexts, which makes the problem non-convex.

Before implementing the negative sampling, we need to assign the window size that defines how many neighbors of words to consider. For example, let's assume that we want to select four neighboring words of 'princess'. According to the word embedding network, there are only a few words located close in the latent Euclidean space, e.g., 'horse', 'money', 'king', 'queen', 'princess', 'prince', 'palace', 'flowers', and '...'. Since we want to select four neighboring words, we set window size as 2 so that two words from each side centering 'princess' can be considered. Therefore, we define 'king', 'queen', 'prince', and 'palace' as near words for the word 'princess'. On the other hand, we can sample 20 negative words that are not included in these near word sets. By repeating this process, it trains the words' network that reflects the context. In this way, we could obtain the word embedding network with 256 dimensions. Using the trained `word2vec` network, we selected ten words from the abstract nouns that were near to each abstract keyword.
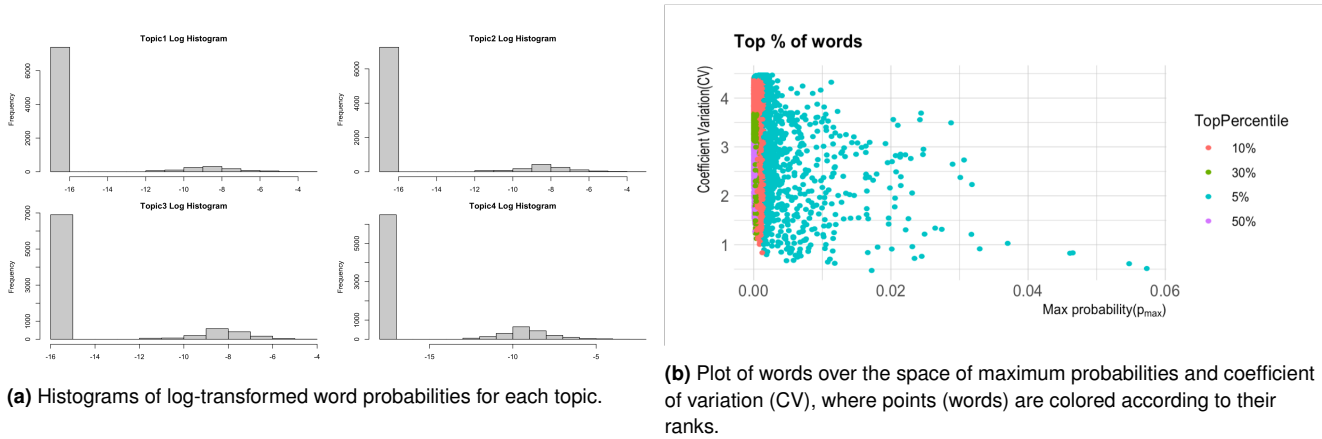
Table 1 shows the summary of corpus construction results with 9,643 words from 15,015 documents. We further filtered out noise words including single alphabet, numbers, and other words that are not meaningful, e.g., 'p.001', 'p.05', 'n1427', 'l.', and 'ie'. Finally, to obtain more meaningful topics, we removed common words like 'data', 'analysis', 'fact', and 'disease'. The full list of filtered words can be found in Appendix.

	Documents	Keywords in Corpus
COVID-19	15,015	9,643

Table 1. Summary of the corpus construction results.

COVID-19 Topic Analysis

To implement BTM, we set the number of topic as 20. For the hyper-parameters, we assigned $\alpha = 3$ and $\beta = 0.01$. Since our main goal is to visualize the topic relationships, we empirically searched and determined the hyper-parameters. The posterior distribution of topic-word was estimated using the Gibbs sampler, where we generated samples with 50,000 iterations after the 20,000 burn-in iterations, and then implemented thinning for every 100th iteration. Table 2 shows the structure of topic-word distribution indicating degree of relatedness of words to a specific topic.

(a) Histograms of log-transformed word probabilities for each topic.

(b) Plot of words over the space of maximum probabilities and coefficient of variation (CV), where points (words) are colored according to their ranks.

Figure 4. (a) Histograms of log-transformed word probabilities for Topics 1, 2, 3, and 4, which show bimodal shapes. The mode on the left correspond to the words that had low possibilities of association with the topic. The mode on the right corresponds to the words with sufficient information to capture meaning of the topic. (b) The top 5% of words were colored blue based on coefficient variation and maximum probability.

Words	Topic1	Topic2	...	Topic 20
nutrient	8.4E-08	9.8E-08	...	9.1E-08
relationship	1.0E-03	1.5E-03	...	2.7E-03
inflammation	1.4E-04	5.8E-04	...	6.3E-04
...	...	...	...	...

Table 2. The topic-word matrix.

Meaningful Top Words

In each topic-word distribution obtained from BTM, words with high probabilities characterize the topic. Figure 4a shows histograms of log-transformed probabilities for each of Topics 1 - 4 (histograms for the other topics can be found in Appendix) and it indicates bimodal topic-word distribution. Specifically, the mode on the left corresponds to the words that had low possibilities of belonging to a specific topic, whereas the mode in the center corresponds to the words that had high probabilities enough to characterize meaning of the topic. Therefore, to reduce noises, it might be desirable to estimate topic relationships using only the words corresponding to the mode in the center, rather than using all the words. Based on these histograms, the minimum cutoff value that defined the tail of a normal distribution was ranged between -11 and -12. We calculated the number of words whose log-scaled probabilities were above -11 to -12 to specify the minimum number of words for estimating the topic network, and we found that more than 1,000 words are needed to properly represent topics. Based on this rationale, we decided to use at least 1,000 words to estimate positions of topics on the latent space based on positions of words.

In order to extract meaningful words that can discriminate characteristics of topics, we chose words with probabilities varying from topic to topic and with relatively high maximum probabilities. Figure 4b shows a plot of the relationship between max probability and coefficient variation. Based on this rationale, we selected words using both max probability and coefficient variation. In this study, rather than using a fixed number of words, we investigated relationships among topics identified with different numbers of words, which were determined using the two criteria described above. Specifically, we obtained

multiple matrices corresponding to the top 60% to 40% of words determined based on the two criteria. The numbers of words corresponding to the 60-th and 40-th percentiles were 2,648 and 1,095, respectively. At the end of the day, we used 21 sets of matrices, $\mathbf{X}_i(i = 1, 2, \cdots, 21)$, as the LSIRM input data, where their dimensions were ranged from $2,648 \times 20$ to $1,095 \times 20$.

Latent Space Item Response Model for Topics' Latent Positions

From the previous steps, we obtained the 21 sets of matrices $\mathbf{X}_i$ and we considered 20 items for all the matrix sets. To estimate topics' latent positions $\mathbf{V} = \{\mathbf{v_i}\}$, MCMC was implemented as described in Step 3 in Methodology section. The MCMC was run 55,000 iterations and the first 5,000 iterations were discarded as a burned-in process. Then, from the remaining 50,000 iterations, we collected 10,000 samples using thinning of 5. To visualize relationships among topics, we used two-dimensional Euclidean space. Additionally, we set 0.28 for $\boldsymbol{\beta}$ jumping rule, 1 for $\boldsymbol{\theta}$ jumping rule, and 0.06 for $\mathbf{w}_j$ and $\mathbf{z}_i$ jumping rules. Here, we fixed prior $\boldsymbol{\beta}$ and let $\boldsymbol{\theta}$ follow N(0,1). We set $a_\sigma = b_\sigma = 0.001$. LSIRM takes each matrix $\mathbf{X}_i$ obtained from the previous step as input and provides the $\mathbf{A}_i$ matrix as output after the Procrustes-Matching within the model. Since we calculate topics' distance on the 2-dimensional Euclidean space, $\mathbf{A}_i$ is of dimension 20×2.

Baseline Topic Network As we mentioned in Step 4, we visualized relationships among topics using the baseline matrix $\mathbf{A}_{\max}$ chosen so that we can compare topics' latent positions without having identifiability issues from the invariance property. From $\mathbf{A}_i$, we calculated the distance between origin and each topic's coordinates. The closer distance of a topic position from the origin indicates the weaker dependency with other topics. Figure 5a showed that dependency structure among topics starts to be built up from $\mathbf{A}_{47\%}$. Based on this rationale, we chose $\mathbf{A}_{\mathbf{47}\%}$ as the baseline matrix $\mathbf{A}_{\max}$. With this baseline matrix $\mathbf{A}_{47\%}$, we implemented the Procrustes matching to align the direction of the topic's latent positions from each matrix $\mathbf{A}_i$. Using

(a) Distance Plot of A_i

(b) Baseline topic network: $A_{47\%}$
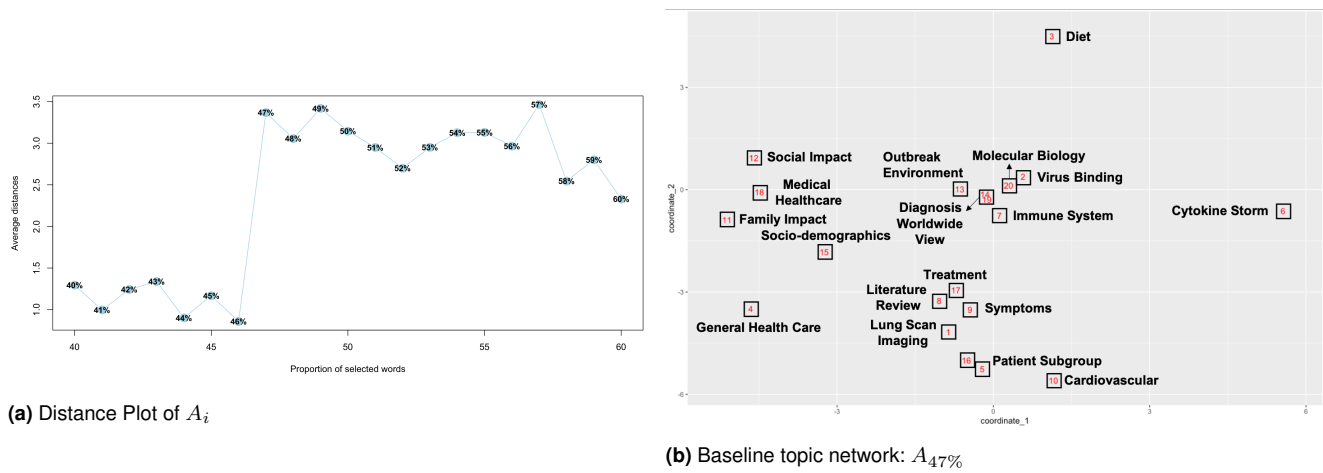
Figure 5. (a) The topic closer to the origin is more likely to have weaker dependency with other topics. It shows that dependency structure among topics starts to be built up from $\mathbf{A}_{47\%}$. (b) The plot of $\mathbf{B}_{47\%}$ shows the topics' latent positions.

this process, we could obtain the $\mathbf{A}_i^*$ matrix matched to the baseline matrix $\mathbf{A}_{47\%}$.

Topic naming We named the identified topics based on top ranking words using the $\mathbf{A}_{47\%}^*$ matrix. This is because the baseline matrix $\mathbf{A}_{47\%}^*$ has the most substantial dependency structure comparing other $\mathbf{A}_i^*$ matrix containing the words that characterize topics nicely. Table 3 shows the name of each topic determined based on their associated words. The top 30 words for each topic identified using $\mathbf{A}_{47\%}^*$ can be found in Appendix.

The first topic is about lung cancer imaging and includes words such as 'chest', 'tomography', and 'scan'. The second topic discusses binding mechanism of COVID-19, which includes words like 'dock(ing)', 'bind(ing)', and 'spike'. The third topic is related to dietary behaviors in the context of COVID-19 and includes words like 'vitamin', 'nutrient', 'carbon', and 'diet(ary)'. The fourth topic discusses various issues related to general health care such as 'healthcare', 'surgery', 'telemedicine', 'staff', and 'equipment'. The fifth topic is related to population subgroups with potentially lower and higher risks, e.g., 'male', 'older', and 'comobidity'. The sixth topic is related to the cytokine storm, which occurs due to over-activation of immune system and can result in severe damage across multiple organs of COVID-19 patients (Ragab et al. 2020). This topic includes words like 'cytokine', 'storm', 'inflammation', and 'IL-6' The seventh topic discusses the immune system more in general context, e.g., various immune cell subsets such as 'lymphocyte', 'neutrophil', and 'CD8 T cell'. The eighth topic is relevant to the literature review, e.g., 'database', 'review', 'literature', 'MEDLINE', 'meta-analysis', 'PubMed', and 'Scopus'. The ninth topic describes various types of COVID-19 symptoms such as 'anosmia' and 'hyposmia', i.e., loss of 'smell' and 'taste'. The tenth topic includes various words related to the cardiovascular system, including 'cardiovascular', 'myocardiac', 'thrombosis', 'stroke', and 'coagulation'. The eleventh topic discusses various impact on family such as 'mental', 'children', and 'school'. The twelfth topic discusses more general social impact, including 'sector' and 'vulnerable'. The thirteenth topic describes the environment where the COVID-19 outbreak occurred,

Topic	Name
1	Lung Scan Imaging
2	Virus Binding
3	Diet
4	General Health Care
5	Patient Subgroup
6	Cytokines Storm
7	Immune System
8	Literature Review
9	Symptoms
10	Cardiovascular
11	Family Impact
12	Social Impact
13	Outbreak Environment
14	Diagnosis
15	Socio-demographics
16	Unknown (Hard to characterize)
17	Treatment
18	Medical Healthcare
19	Worldwide View
20	Molecular Biology

Table 3. Topic names based on $A_{47\%}^*$ matrix.

e.g., 'Wuhan', 'Hubei', 'Korea', 'province', 'temperature', and 'pollution'. The fourteenth topic is related to the diagnosis of COVID-19 such as 'swab', 'IgG', 'PCR', 'IgM', 'immunoassay', and 'ELISA'. The fifteenth and sixteenth topics are somewhat vague to characterize but the former seems to be relevant to socio-demographics. The seventeenth topic is about the treatment of COVID-19 and includes words like 'Hydroxychloroquine', 'Azithromycin', and 'Tocilizumab'. The eighteenth topic discusses more specific medical healthcare such as 'mask', 'equipment', 'PPE', 'surgical', 'supply', 'glove', and 'shield'. The nineteenth topic generally describes what happens worldwide (e.g., 'worldwide' and 'globe'). Finally, the twentieth topic is specific to molecular biology related to COVID-19, which includes words such as 'genome', 'epitope', 'amino acid', 'phylogenetics', and 'nucleotide'.

Oblique Rotation To improve interpretation of the identified latent space (e.g., finding the meaning of topics'

transition based on the X-axis or Y-axis), we rotated the original latent space so that the axes better encompass topics. We applied `oblim` rotation to the estimated topic position matrix $\mathbf{A}^*_{47\%}$ using the R package GPArotation (Bernaards and Jennrich 2005), and obtained matrix $\mathbf{B_i}$ with the rotation matrix $\mathbf{R}$. In the same way, we rotated the other estimated topic position matrix $\mathbf{A}^*_i$ resulting in $\mathbf{B_i}$ for $i = 40\%, \cdots, 60\%$. Figure 5b shows $\mathbf{B_{47\%}}$ representing the topics' latent positions. In this figure, the topics in the center include 'Outbreak Environment', 'Diagnosis', 'Worldwide View', and 'Immune System'. In contrast, 'Molecular Biology' corresponds to the case of weak dependency with other topics. There are two possibilities that lead to low dependency. First, it is possible that there were a small number of words that can distinguish their topics' characteristics from the other topics. Second, it is also possible that most of the words were common words shared with other topics. To eliminate visualization bias due to selection of the number of words, we tracked the topics' latent positions by using different input matrices $\mathbf{B_i}$.

## Trajectory Plot for Visualizing Relationships among Topics

In this section, using the results obtained with the procedure described above, we interrogated what kind of topics have been extensively studied in the biomedical literature on COVID-19. In addition, we also studied how those topics were related to each other, based on their closeness in the sense of latent positions. We also partially clustered the topics based on their relationships using the quadrants. This allows us to check which studies about COVID19 are relevant to each other and could be integrated. Figure 6 displays the trajectory plot and it shows how topics were positioned on the latent space and how these topics make transition. More specifically, the direction of arrows refers to how topics' coordinates changed as a function of numbers of words, where each arrow moves from $\mathbf{B_{60\%}}$ to $\mathbf{B_{40\%}}$ as the number of words decreases.

Overall Relationship among Topics According to Figure 6, we observed two meaningful distinct groups: One group independent of other topics and the other with strong dependency with other topics. First, the topics 'Outbreak Environment', 'Diagnosis', and 'Worldwide' (Topics 13, 14, and 19, respectively) were located in the center of the plot. This indicates that no matter how many words were used to estimate the topics' latent positions, those topics remained as general topics and have shares many words with other topics. This makes nice sense given the fact that many articles have mentioned the outbreak of COVID-19, the diagnosis of COIVD-19, and how the world treats or considers this notorious pandemics. For example, among the 137K publications mentioning "COVID-19" in the PubMed database, more than 76,000 and 35,000 publications also mentioned "outbreak" and "diagnosis", respectively. Second, the topics 'Social Impact', 'Family Impact', 'Cardiovascular', and 'Cytokine Storm' (Topics 12, 11, 10, and 6, respectively) were located away from the center in the plot, which implies their dependency structures with other topics. These topics usually stay on the boundary of the plot regardless of the number of words because they

consist mainly of unique words. Finally, the topics like 'Socio-demographics', 'General Health Care', and 'Virus Binding' (Topics 15, 4, and 2, respectively) start from the origin, stay away from the origin for a while, and then return to the origin. This implies that it could not maintain the nature of the topic when less words are considered, and it is likely that those topics are either ongoing research or burgeoning topic that is not studied enough yet.

Partial Relationship based on Quadrant In the previous section, we interpreted the topics' meaning based on their latent positions. In this section, we now make interpretation using subsets of topics divided by the axes. Since we implemented the oblique rotation that maximizes each axis' distinct meaning, we can render meaning to an axis. Figure 6 indicates that there are five topic clusters. First, the center cluster denoted as (A) in Figure 6 are about the outbreak of COVID-19: The world view of COVID-19, diagnosis of cover 19, and the environment of outbreak COVID-19. These topics are positioned in the center because they share the common words with the other topics, i.e., most literature mention the outbreak of COVID-19. Second, the topics located on the left side of the plot (cluster (B) in Figure 6) are related to the impact of COVID-19 at the society level. For instance, the impact of COVID-19 on health care centers, including hospitals, are visible phenomena of COVID-19. These subjects pertain to 'Social Impact (12)', 'Medical Healthcare (18)', 'Family Impact (11),' 'Socio-demographics (15)', and 'General Health Care (4)'. Third, the next cluster, called (C), is relevant to the strategies to study COVID-19. For instance, this cluster includes the literature review of COVID-19, patients' lung scan image, symptoms, and treatment. Fourth, the topics corresponding to the cluster (D) are related to what happens inside our body in response to COVID-19, e.g., cytokine storm, immune system responding to the COVID-19 infection, and cardiovascular impacts of COVID-19. Finally, the cluster (E) is related to molecular changes upon the COVID-19 infection, e.g., binding mechanism of SARS-CoV-2. We note that, aforementioned, we can also interpret each axis and x axis can be interpreted as spectrum from macro (more negative) to micro perspectives (more positive). In summary, we identified five main groups of the COVID-19 literature: The outbreak of COVID-19, lifestyle after COVID-19 attack, COVID-19 study strategies, impacts of COVID-19 effect on the body, and molecular changes by COVID-19 infection counterclockwise.

Counter Clockwise Relationship Here we can derive another insight from the locations of clusters. Specifically, from the cluster (A) to the cluster (E), we can observe transition from macro perspectives to micro perspectives clockwise. Specifically, this flow starts with the center cluster (A) related to the occurrence of COVID-19, followed by studies of the social impact of COVID-19 (cluster (B)) and then patient-level investigation (cluster (C); lunch scanning and symptoms). This flow ends with the clusters (D) and (E), which are related to micro-level events, e.g., how SARS-CoV-2 binding occurs and how the immune system responds to upon the COVID-19 infection.
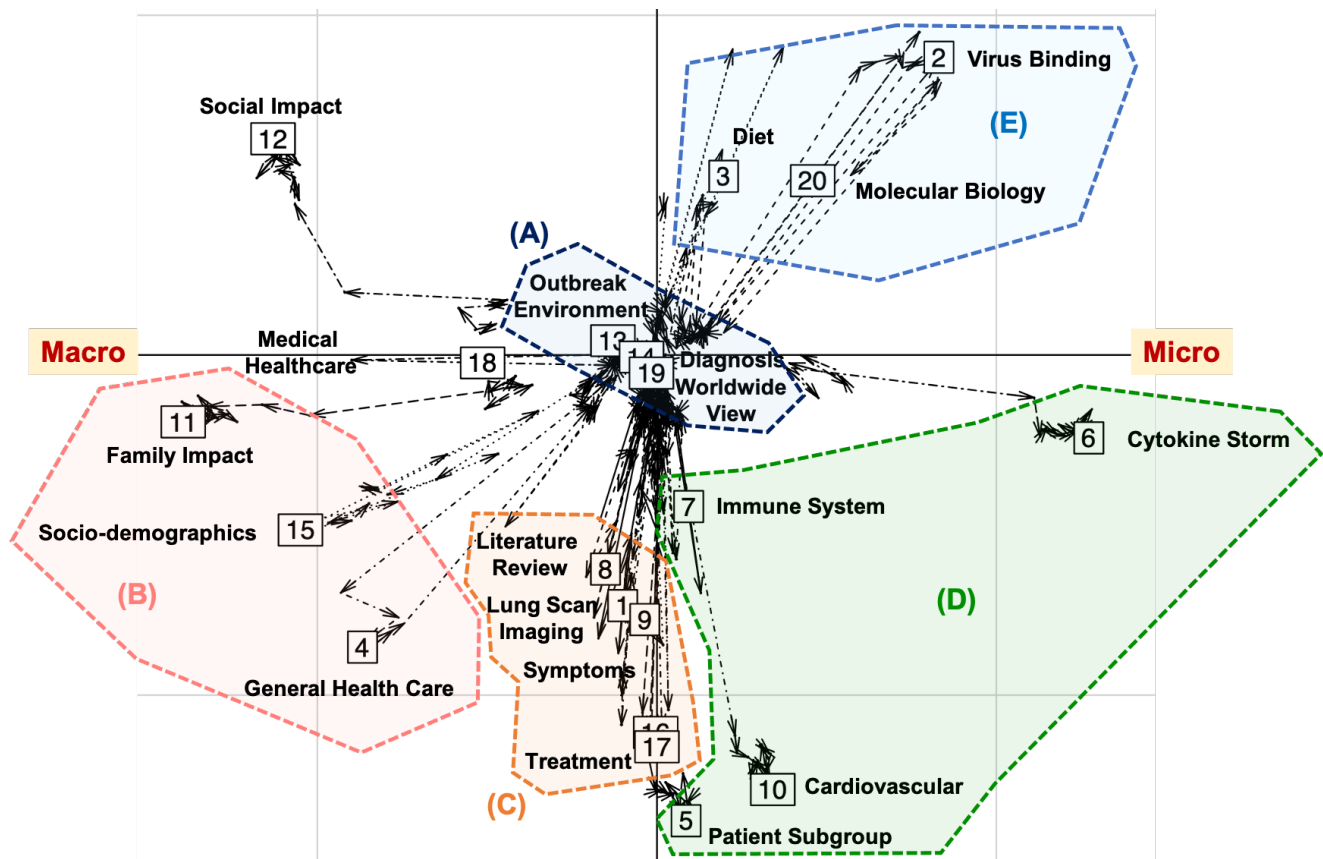
Figure 6. Trajectory plot with five topic clusters (A) - (E). (A) Outbreak of COVID-19. (B) Lifestyle after COVID19-attack. (C) Strategies to study COVID-19. (D) Response of our body upon the COVID-19 infection. (E) Molecular phenomena associated with the COVID-19 infection.

Discussion

In this manuscript, we proposed a novel analysis framework to visualize relationships among topics based on the text mining results. It allows enhancing our understanding of COVID-19 knowledge reported in the biomedical literature, by evaluating topics' networks through their latent positions estimated based on topic sharing patterns. The proposed approach overcame limitations of existing approaches, especially discrete and static visualization of relationship among topics. First, because we position topics on a latent space, relationships among topics can be intuitively investigated and also easy to interpret. Second, our method allows deeper understanding of the network among topics and capturing their dynamics with a continuous representation by tracking the change of topics' relationships. To the authors' best knowledge, this is the first attempt to integrate the biterm model and the latent space item response model within a unified framework. The application of our method to COVID-19 literature indicated that there are five main subjects in the COVID-19 biomedical literature. The proposed framework can still be further improved in several ways, especially by allowing word-level inference, i.e., extraction of meaningful words that characterize each topic. Although distance between each topic and relevant words is taken into account in our model when estimating topics' latent positions, simultaneous representation and visualization of words are still not embedded in the current framework. We believe

that adopting a variable selection procedure to determine key words can potentially address this issue and this will be an interesting future research avenue.

Acknowledgements

References

Amado A, Cortez P, Rita P and Moro S (2018) Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics* 24(1): 1–7.

Andrews K, Kienreich W, Sabol V, Becker J, Droschl G, Kappe F, Granitzer M, Auer P and Tochtermann K (2002) The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization* 1(3-4): 166–181.

Bernaards CA and Jennrich RI (2005) Gradient projection algorithms and software for arbitrary rotation criteria in factor

analysis. *Educational and Psychological Measurement* 65(5): 676–696.

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.

Borg I and Groenen PJ (2005) *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Cao N, Sun J, Lin YR, Gotz D, Liu S and Qu H (2010) Facetatlas: Multifaceted visualization for rich text corpora. *IEEE transactions on visualization and computer graphics* 16(6): 1172–1181.

Chen GB and Kao HY (2017) Word co-occurrence augmented topic model in short text. *Intelligent Data Analysis* 21(S1): S55–S70.

Chen Q, Allot A and Lu Z (2020) Keep up with the latest coronavirus research. *Nature* 579(7798): 193. DOI:10.1038/d41586-020-00694-1. URL https://www.ncbi.nlm.nih.gov/pubmed/32157233.

Chowdhury GG (2010) *Introduction to modern information retrieval*. Facet publishing.

Deerwester S, Dumais ST, Furnas GW, Landauer TK and Harshman R (1990) Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6): 391–407.

Dong E, Du H and Gardner L (2020) An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* 20(5): 533–534.

Goldberg Y and Levy O (2014) word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* .

Gretarsson B, O'donovan J, Bostandjiev S, Höllerer T, Asuncion A, Newman D and Smyth P (2012) Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(2): 1–26.

Griffiths TL and Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(1): 5228–5235.

Gruber A, Weiss Y and Rosen-Zvi M (2007) Hidden topic markov models. In: *Artificial intelligence and statistics*. PMLR, pp. 163–170.

Hoff PD, Raftery AE and Handcock MS (2002) Latent space approaches to social network analysis. *Journal of the american Statistical association* 97(460): 1090–1098.

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 42(1): 177–196.

Jennrich RI (2002) A simple general method for oblique rotation. *Psychometrika* 67(1): 7–19.

Jeon M, Jin I, Schweinberger M and Baugh S (2021) Mapping unobserved item–respondent interactions: A latent space item response model with interaction map. *Psychometrika* DOI: https://doi.org/10.1007/s11336-021-09762-5.

Jiang S, Qian X, Shen J, Fu Y and Mei T (2015) Author topic model-based collaborative filtering for personalized poi recommendations. *IEEE transactions on multimedia* 17(6): 907–918.

Lafferty JD and Blei DM (2006) Correlated topic models. *Advances in neural information processing systems* 18: 147–154.

Lakkaraju H, Bhattacharya I and Bhattacharyya C (2012) Dynamic multi-relational chinese restaurant process for analyzing influences on users in social media. In: *2012 IEEE 12th International Conference on Data Mining*. IEEE, pp. 389–398.

Li W, Feng Y, Li D and Yu Z (2016) Micro-blog topic detection method based on btm topic model and k-means clustering algorithm. *Automatic Control and Computer Sciences* 50(4): 271–277.

Liu JS (1994) The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* 89(427): 958–966.

Liu L, Tang L, Dong W, Yao S and Zhou W (2016) An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5(1): 1608.

Mei Q, Cai D, Zhang D and Zhai C (2008) Topic modeling with network regularization. In: *Proceedings of the 17th international conference on World Wide Web*. pp. 101–110.

Mikolov T, Chen K, Corrado G and Dean J (2013a) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mikolov T, Sutskever I, Chen K, Corrado G and Dean J (2013b) Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* .

Newman D, Baldwin T, Cavedon L, Huang E, Karimi S, Martinez D, Scholer F and Zobel J (2010) Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2-3): 169–175.

Nigam K, McCallum AK, Thrun S and Mitchell T (2000) Text classification from labeled and unlabeled documents using em. *Machine learning* 39(2): 103–134.

Ragab D, Salah Eldin H, Taeimah M, Khattab R and Salem R (2020) The COVID-19 cytokine storm; what we know so far. *Frontiers in Immunology* 11: 1446.

Shortreed S, Handcock MS and Hoff P (2006) Positional estimation within a latent space model for networks. *Methodology* 2(1): 24–33.

Sievert C and Shirley K (2014) Ldavis: A method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. pp. 63–70.

Yan X, Guo J, Lan Y and Cheng X (2013) A biterm topic model for short texts. In: *Proceedings of the 22nd international conference on World Wide Web*. pp. 1445–1456.

Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H and Li X (2011) Comparing twitter and traditional media using topic models. In: *European conference on information retrieval*. Springer, pp. 338–349.