# NEWS DATA VISUALIZATION INTERFACE DEVELOPMENT USING NMF ALGORITHM

A Thesis
Presented to
The Academic Faculty

By

Byeongsoo Ahn

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Computational Science and Engineering
College of Computing

Georgia Institute of Technology

May 2022

**NEWS DATA VISUALIZATION INTERFACE DEVELOPMENT USING NMF ALGORITHM**

Thesis committee:

Dr. Haesun Park
School of Computational Science and Engineering
*Georgia Institute of Technology*

Dr. Ling Liu
School of Computer Science
*Georgia Institute of Technology*

Dr. Sungha Kang
School of Mathematics
*Georgia Institute of Technology*

Date approved: April 29, 2022

No pain, No gain.

*Benjamin Franklin*

For Republic of Korea, my proud homeland

## ACKNOWLEDGMENTS

I first want to thank my advisor, Dr. Haesun Park for her all guidance and support during my masters years. I also like to thank Dr. Liu and Dr. Kang, who are my committee members, for advising and guiding me. Moreover, without the professors who directly and indirectly influenced me at Georgia Tech, this thesis would not have been completed.

I would also like to thank Dongjin Choi, my lab mate and mentor. He always led me whenever I had trouble for the study, and was a great help in writing this thesis. He also provided me with a mental support.

I would like to thank Robin, the most precious relationship between life in America and my eternal partner. I was able to complete this thesis thanks to her always by my side, supporting me emotionally and protecting me.

I would like to express my gratitude to my precious classmate Youngjoon, who came to the United States to study at the same time and gave me strength whenever I was having a hard time.

I would also like to express my gratitude to my high school friends in Korea, my classmates from the Korean Military Academy, tennis club members, undergraduate friends, and all other friends.

I cannot thank my parents and my sister enough for their endless love and support. Thank you for always believing in me, praying for me, and loving me unconditionally.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ACRONYMS

**LDA** Latent Dirichlet Allocation

**NLP** Natural language processing

**NMF** Nonnegative Matrix Factorization

**PLSA** Probabilistic Latent Semantic Analysis

**SVD** Singular Value Decomposition

**UI** User Interface

## SUMMARY

News data is a super large-scale dataset. It covers a wide range of topics ranging from heavy topics such as politics and society to beauty and entertainment, relatively light topics. At the same time, it is also the most accessible source of information for the general public to obtain information.

Thus, how is this large amount of data used by the general public being utilized? Currently, services provided by news platforms are just full article searches and related news recommendations. It uses only a fraction of the vast news dataset, and there is still a lack of systems to fully utilize and analyze it.

As mentioned above, news datasets which contain a wide range of topics and super-large scales of data, record everything that happened in the past and present, so analyzing and visualizing them can track how trends in real-world change over time and even discover what the topics of the large dataset are without reading the full text through topic modeling. For this objective, in this thesis, we propose a novel interactive visualization interface for the news data based on NMF to analyze, visualize, and utilize datasets more practically than simply searching the articles.

Through this thesis, We first suggest a prototype visual interface that visualizes the topic modeling results of the news dataset over time. This interface is a novel approach that connects news data, visual interfaces, and topic modeling at once. Then, the most suitable method for the interactive visual interface is presented by comparing various topic modeling methods. Finally, we present use cases on how this study can be used practically and present their applicability in various fields.

# CHAPTER 1
# INTRODUCTION

News data is one of the most accessible sources for the public to gain information and knowledge. In particular, due to the nature of news data that deals with various topics and accumulates more than a thousand articles per day, the scale of data continuously increases. Therefore, the systems which analyze it are essential to handle and fully understand such information.

Compared to the potential power of the vast dataset, current users' use of news data is limited to news searches, and services provided by current platforms are also limited to providing search results and recommending articles. However, it uses only a fraction of the information obtained from the dataset. Users can analyze the impact of a specific event through the increase or decrease in the number of articles related to that event or analyze detailed topic keywords that are not well revealed by major categories such as politics, society, and sports. There are many other ways to use it, but the systems and interfaces that support it are insufficient.

This thesis focuses on topic modeling and trend visualization of news data to help users understand it. To this end, NMF is selected and applied as the back-end method through the comparison experiment of the related work and the processing speed comparison experiment performed in this study, and a novel visual interface is developed to illustrate the result.

NMF is *Nonnegative Matrix Factorization*, which is a main method used for clustering and topic modeling in this study. The mathematical principle of NMF is straightforward. When we have a matrix $A \in \mathbb{R}^{m \times n}$, factorize it into two nonnegative matrices $W$ and $H$ such that

$$A \approx WH \quad \text{s.t. } W, H \geq 0$$

It is a straightforward principle, but there are numerous applicable fields, including image processing [1], subsystem identification [2], cancer class discovery [3], and text data mining [4], and this study focuses on text data mining among them. Due to the characteristic of the unsupervised method, target text data such as papers and articles do not require additional labeling or annotations for applying this algorithm. Therefore this method can save time and cost for the task. In this study, a matrix consisting of news data articles and the term included therein is constructed. Based on NMF results, finding the top-N-keyword is performed by analyzing $W$, and document clustering is performed using $H$ to determine which topic each article belongs to, which is described in more detail in Chapter 3.

Even if algorithms or methods running on the back-end perform well, it can be somewhat complicated for the general public, the primary users of news data. In fact, there has been continuous research on data mining using news data [5] [6] [7], and in particular, there is a study in which topic modeling has been performed by applying the NMF method to news data [8]. Nevertheless, it is scarce for studies to connect all three: news data, NMF algorithm, and Visualization. To this end, we develop a more intuitive and user-friendly visual interface that can show information and trends in news data and topic modeling results at once. Users can explore the topic modeling result and analyze the trends over time through the visual interface and check the whole text of the article. Significantly, the trend is effectively shown by visualizing the increase or decrease in the number of articles corresponding to each topic in two ways: 3d bar plot and stack area plot. Through this, the user can analyze a specific topic by checking a graph that shows dramatic changes at a specific point in time, explore keywords corresponding to the topic, and obtain additional information through articles belonging to the topic.

In this thesis, we suggest three contributions to this study. First of all, we developed a novel prototype visual interface that combines three elements: (1) news data, (2) topic modeling, and (3) Visual Interface. This visual interface also includes two effective topic visualization features, a stack area plot and a 3-D plot to realize better data analysis. The

second is the back-end method. Among the various methods, we demonstrated the NMF's superior accuracy and processing speed performance and applied it to the interactive system. The third is practicality. Through a use case study, this thesis suggests what scenarios this study can use in practice.

The rest of the paper is as follows. Chapter 2 introduces related works on topic modeling and visual analysis of text data over time. Chapter 3 explains the main algorithm, NMF, and other efficient computation methods used in this study. Chapter 4 presents the design structures of the visual interface and its various features. Chapter 5 is a section that explains the performance of the method and the interface. It also includes user cases, which can be an example of actual use. In the end, chapter 6 presents the conclusion of this thesis.

# CHAPTER 2
# RELATED WORK

This chapter describes related work on the three most important concepts in this study. As can be observed from the below, each concept is being actively researched, and research in which the two concepts are correlated, such as topic modeling and news data, topic modeling and visualization, and news data and visualization, is being actively conducted. However, not many studies include all of them, so this point motivates us to conduct this study.

## 2.1 Topic modeling method

Research to interpret documents or text data mathematically and computationally has been steadily conducted in the past. In particular, attempts to interpret semantic relationships through probabilistic models were made [9] [10]. Probabilistic topic modeling was introduced to take advantage of the increasing scale of data. This concept discovers themes based on the statistical information of documents and words contained in the documents [11]. Since topic modeling was introduced, various models have been proposed to implement it.

The first approach is utilizing Singular Value Decomposition (SVD). Since the method using decomposition of term-document matrices with SVD was suggested [12], many studies relied on the method. But the studies with SVD has limitations : (1) it needs the assumption that documents have just one topic, and (2) It is not able to find topic vector themselves but only finds the span of the topic vectors [13]

Another model is Probabilistic Latent Semantic Analysis (PLSA) [14]. It adopted a probabilistic approach to the problem. It also has some limitations, such as overfitting.

On the other hand, one of the recently used models is LDA. The idea of LDA starts

that the documents are a mixture of latent topics, and these topics are represented as a distribution of words [15]. In particular, LDA is one of the most popular models, so it is not just applied to text data analysis but has been utilized in various domains. For example, a model using LDA is introduced and utilized for health care [16] [17], geography [18] [19], and even music [20].

The last model is NMF, first suggested by Paatero and Tapper [21]. It is used as the primary method in this study, and it shows excellent results as well as LDA, so this method is also adopted in various fields. In the political field, it was used to analyze the agenda [22], and in bioinformatics, it was used to observe cancer-related data [23] and also used for computational biology analysis [24]. In addition to this, it has demonstrated its excellence in computer vision [25] and other fields.

#### 2.1.1 Evaluation over methods

As various methods, including the four methods introduced above, have been introduced, research to compare their performance has been steadily progressing.

In a study comparing the topic coherence performance of two methods, NMF and LDA, NMF showed a more coherent topic descriptor overall [26].

In a study comparing three topic models (NMF, LDA, SVD), an experiment was also conducted to evaluate coherence according to the increasing number of topic numbers [27].

### 2.2 News data analysis

News data used in this study also been also used in various studies. Like this study, semantic analysis, such as topic modeling, is also a major field. In addition to English news, research using domestic news data such as Swedish news [8], Turkish news[28], and Korean news[29] is also being actively conducted. Semantic analysis through articles is also a major research field [30] [5] [31] [32] [33].

## 2.3 Visual interface

In the field of Natural language processing (NLP), visualization is a well-known area in terms of showing research results more effectively and helping readers understand. VisIrr presented an interactive visual system capable of information retrieval and recommendation using a Large document dataset [34], and UTOPIAN presented a system that effectively visualizes the results of topic modeling [35]. Also, Architext presented an interactive system that supports hierarchical analysis of topic modeling [36].

### 2.3.1 Visualization over time

Research to visualize changes over time is also being actively conducted. There is a representative ThemeRiver, which visualizes the thematic changes of document data according to the flow [37] [38], and research has been conducted to provide visualization corresponding to the time specified in SPIRE [39]. In CiteSpace II, a system to analyze and visualize patterns and trends of scientific documents was presented [40].

# CHAPTER 3

# WORKFLOW

As mentioned in Chapter 2, research related to the three concepts, topic modeling, visual interface, and news dataset, has been actively conducted in the past. However, studies such as topic modeling through news data and visualization of its trend over time, which connects the three concepts simultaneously, are rare. One of the possible reasons we considered is that the performance of the existing methods for topic modeling is not adequate to produce meaningful results for a large dataset. The solutions we suggest are (1) reducing the computational cost of back-end computing through pre-filtering of datasets and (2) using the NMF method that shows superior processing speed with high quality of topic modeling when the number of topics is small(˜ 20) compared to other existing methods. In this section, we first define the problem and then describe our algorithm.

### 3.0.1 Problem Formulation

When we have a collection of the total $n$ articles $O = \{o_1, o_2, \cdots, o_n\}$, then after the filtering process, we can define the filtered collection of $m$ articles as $O^* = \{o_1, o_2, \cdots, o_m\}_{(m \leq n)}$. For this target dataset, $O^*$, We define two functions : (1) A function that models $k$ topics defined by the user from the target dataset $O^*$, and (2) A function that assigns each article $o_i$ in $O^*$ into appropriate topics.

### 3.0.2 Data filtering

As mentioned at the beginning of chapter 3, we adopt filtering for the dataset to save computational cost and time. This filtering process receives parameters designated by the user, for example, time range setting, mandatory included keywords, and target categories. These parameters are optional, and if not designated, the entire dataset will be used.

The filtering algorithm is as follows

**Algorithm 1** Algorithm for filtering dataset

**Input:** Article dataset $O$, query term $q$, time range input $t_{from}, t_{to}$, target category list $C$;
**Output:** Filtered Article dataset, $O^*$;

```
for ∀o_i ∈ O do
    if FILTER_1(t_from, t_to, o_i) then
        if FILTER_2(C, o_i) then
            if FILTER_3(q, o_i) then
                return o_i;
return O* ← o_i
```

Note that $FILTER_1(t_{from}, t_{to}, o_i)$, $FILTER_2(C, o_i)$, and $FILTER_3(q, o_I)$ from Algorithm 1 are filtering functions that returns only articles within the time range from $t_{from}$ to $t_{to}$, corresponding to category list $C$, and containing string $q$ in the title or the text of articles.

Through Algorithm 1, the user's target dataset can be set more precisely, and a higher quality topic modeling result can be obtained.

### 3.0.3 Computing topic modeling using NMF

When we have a matrix $A \in \mathbb{R}^{m \times n}$, which is non-negative, and a desired rank $k < min(m,n)$, NMF is to approximate $A$ into two nonnegative factors $W, H$ [41] [4]. It can be re-written as

$$A \approx WH \text{ s.t. } W, H \geq 0, \tag{3.1}$$

Equation 3.1 can be reformulated as the optimization problem :

$$\min_{W,H} f(W,H) \equiv \frac{1}{2}\|A - WH\|_F^2 \text{ s.t. } W, H \geq 0 \tag{3.2}$$

We solve this problem using *Block principal pivoting method* [42] in this study.

In Equation 3.2, $W \in \mathbb{R}^{m \times k}$ is a basis matrix, and $H \in \mathbb{R}^{k \times n}$ is a coefficient matrix [4].

This is the basic formulation of NMF and we can apply it to our study.

First, we can make a nonnegative matrix, $A \in \mathbb{R}^{m \times n}$, from the filtered article set, $O^*$. $A$ is a TF-IDF matrix using the bag-of-word model. $A$ consists of $m$ words from a dictionary made by the article set in its rows and $n$ documents in its columns.

Then, by solving the nonnegative least squares problem shown in Equation 3.2, we get the nonnegative matrices $W$ and $H$. Since $W \in \mathbb{R}^{m \times k}$ is a basis matrix, and $H \in \mathbb{R}^{k \times n}$ is a coefficient matrix, the column vector of each document, $a_j$ , of matrix $A$ can be interpreted as

$$a_j = Wh_j \tag{3.3}$$

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & \ddots & & & \\ a_{3,1} & & & & \\ \vdots & & & & \\ a_{m,1} & & & & \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,k} \\ w_{2,1} & \ddots & \\ w_{3,1} & & \\ \vdots & & \\ w_{m,1} & & \end{bmatrix} \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,n} \\ \vdots & \ddots & & & \\ h_{k,1} & & & & \end{bmatrix} \tag{3.4}$$

Thus, each document, $a_j$ is a linear combination of the columns $w_j$ with coefficients $h_{ij}$. Moreover, if we normalize each column of $W$ when we solve the nonnegative least squares problem, we can interpret $w_j$, the column of $W$, as the score of each word in the topic $j$. Comparing each score, we can extract the top $N$ words representing the topic.

### 3.0.4 Clustering : Assigning each article into each topic

Otherwise, coefficient matrix $H$ consists of $k$ rows, the number of the desired rank, and $n$ columns, which is the number of articles. If we interpret this matrix by column, we can interpret which topic each article is most relevant to. From the interpretation, like subsection 3.0.3, the values of each column vector can be considered as scores, and each article can be clustered to the topic with the highest score.

For example, let us assume we have matrix $A \in \mathbb{R}^{8400\times1102}$ and a desired rank $k = 10$, then we can get a matrix $W \in \mathbb{R}^{8400\times10}$ and $H \in \mathbb{R}^{10\times1102}$ as a result of NMF. If the first column of $H_1$ is,

$$H_1 = \begin{bmatrix} 0.0345 \\ 0.0115 \\ 0.0073 \\ 0.0675 \\ 0 \\ 0 \\ 0 \\ 0.3375 \\ 0 \\ 0 \end{bmatrix}$$

Then the first article of $H$ is clustered into the eighth topic because the eighth score is the highest.

# CHAPTER 4
# VISUAL INTERFACE

We develop a visual interface to fully and effectively display the results of topic modeling and news article trend over time. For these objectives, we summarize three objectives that this interface focuses on as follows.

## 4.1 Visualization Goals

**VG1 Design the interface to be simple and intuitive for the general users.**

The critical thing to consider while designing the visual interface is the users of the interface. This study set the target users as non-experts and the general public. To this end, we try to design it intuitive and straightforward so that anyone can use it right away without additional tutorials or guidelines. In addition, we added essential functions to the interface from the user's point of view. For example, suppose the user wants to search only a specific area rather than the entire dataset. The dataset can be filtered primarily by setting the time range, setting the topic, and entering required keywords. In addition, all visualization components are displayed on one screen so that the user can check the desired information at once.

**VG2 Effectively show that the number of news articles belonging to each topic changes over time.** We also aim to effectively visualize the news dataset topics computed by the NMF method. In particular, rather than showing simple results, the concept of "time" is added to fit the characteristics of news data and how each topic modeling result changes over time. To this end, a stack area plot and a 3D plot are implemented. A stack area plot shows the change in each topic size by month in two dimensions to help intuitive understanding, and in the case of a 3D plot, the bar's height changes by

year. In particular, in the case of a 3d plot, the user can rotate 360 degrees, enabling two-dimensional analysis.

**VG3 Directly illustrate the result of topic modeling computed by the method** In addition, we also aim to effectively display the results of topic modeling through the Visual interface. It is implemented in the Topic Detail Panel, which does not simply list all the results of topic modeling, but shows the top 10 keywords of a topic to show which keywords represent it at a glance.

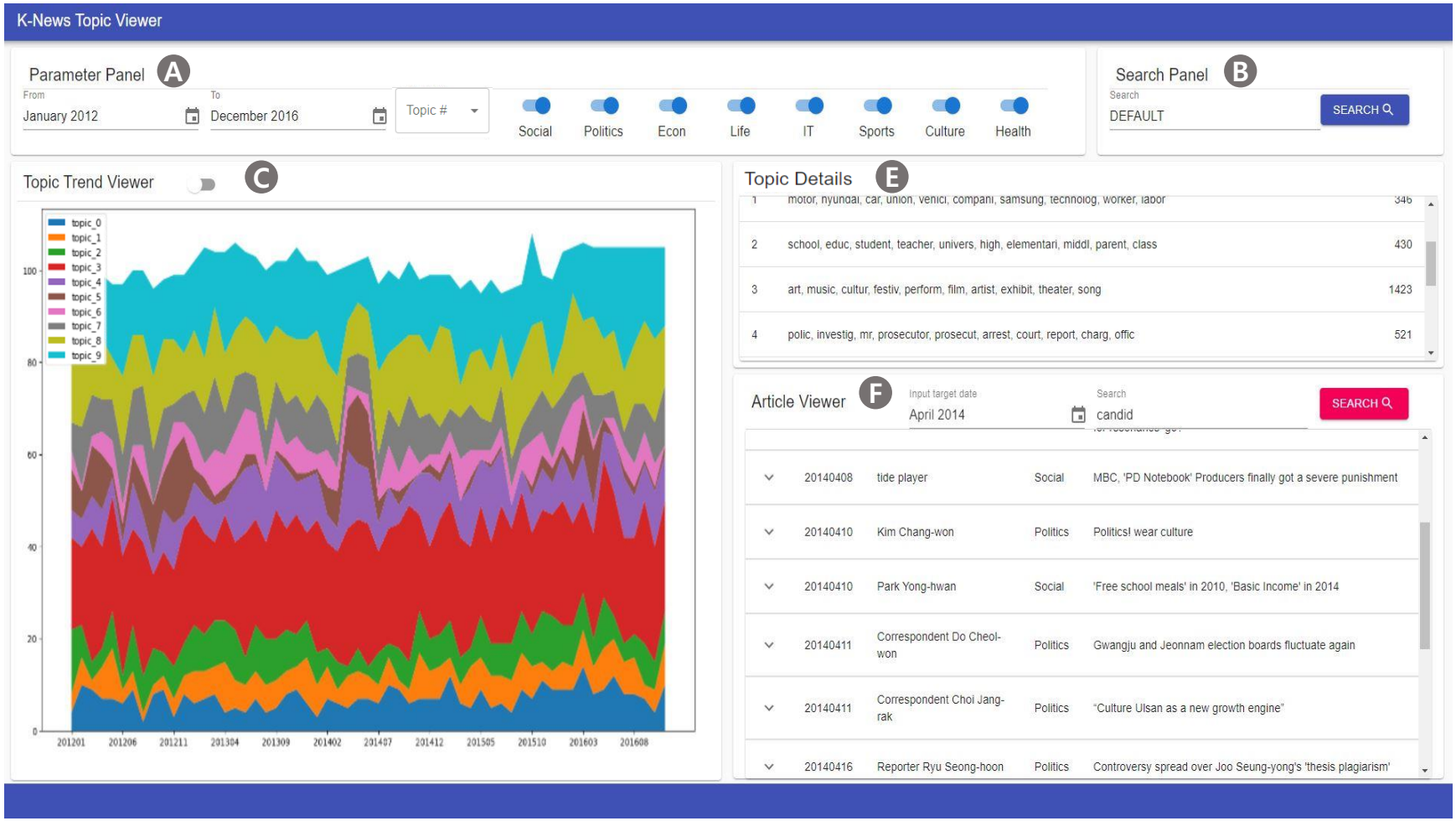


Figure 4.1: UI consists of 5 panels, *Parameter Panel* Ⓐ, *Search Panel* Ⓑ, *Topic Trend Viewer* Ⓒ, *Topic Details* Ⓔ and *Article Viewer* Ⓕ. It is not displayed on the teaser, 3d plot is also displayed on the additional screen.

## 4.2 System Overview

Figure 4.1 illustrates a design overview of the visual interface. The interface is composed of 5 panels and will be explained in detail below.

### 4.2.1 Parameter Panel

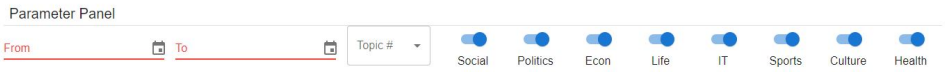


Figure 4.2: Parameter Panel

*Parameter Panel* Ⓐ is the starting point for using the UI. As shown in Figure 4.2, the users can first set the time range that they want to check and choose categories among eight major categories: Social, Politics, Economy, Life, IT/Science, Sports, Culture, and Health. Filtering proceeds according to the parameter values entered into this panel, and the processing speed of the UI using this filtered dataset can be improved. Furthermore, the users can also set the topic number, which is set to the parameter value in the back-end computation.

### 4.2.2 Search Panel

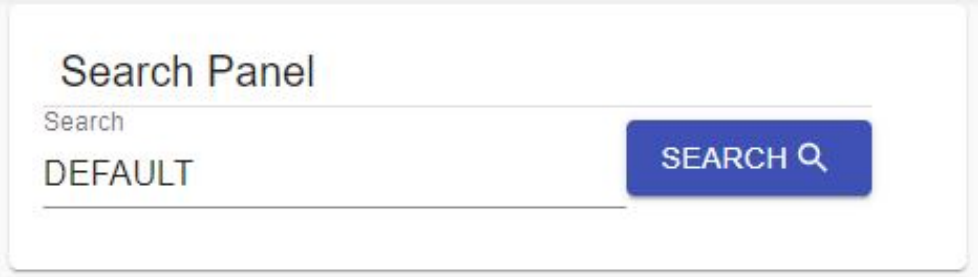


Figure 4.3: Search Panel

*Search Panel* Ⓑ in Figure 4.3 is a panel that allows users to enter a keyword. When the user enters the input and clicks the “Search” button, the input is delivered to the server in

the form of a string, and only articles containing the input are filtered. If the users leave it blank or enter the input 'DEFAULT', the entire dataset will be used.

### 4.2.3 Topic Trend Viewer

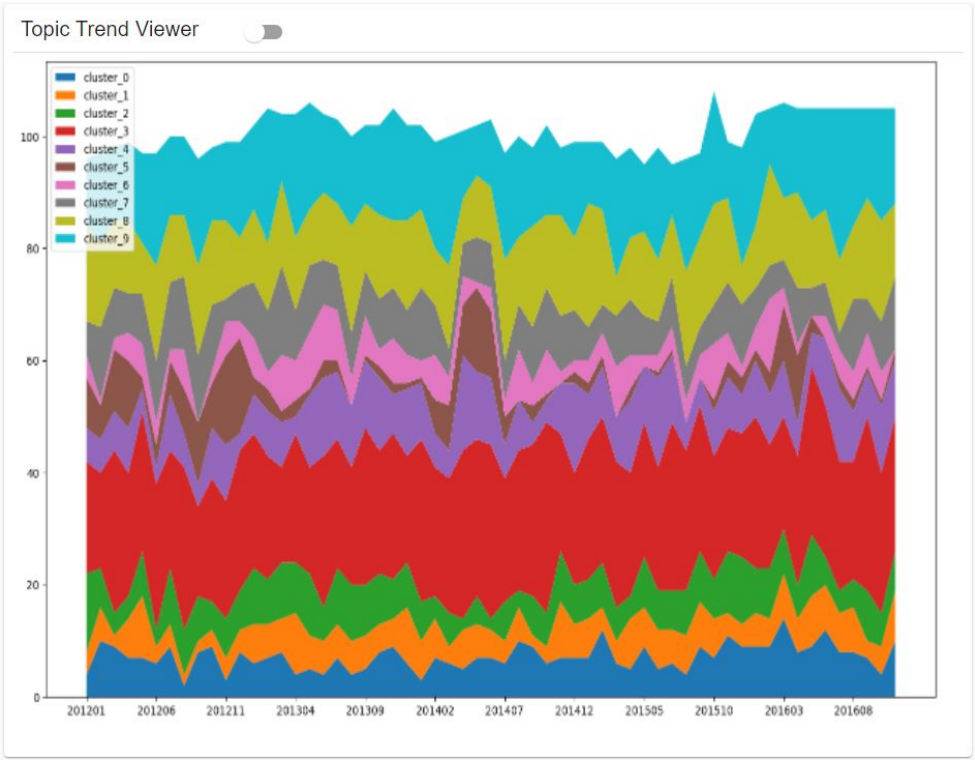


Figure 4.4: Topic Trend Viewer

*Topic Trend Viewer* Ⓒ in Figure 4.4 is a panel that visualizes the topic modeling result computed through the NMF algorithm on the server based on the dataset, which is filtered by parameters and word input given by the user in Ⓐ and Ⓑ. In order to show the 'flow of time', which is an essential attribute of news data, a Stack area plot is shown in the main viewer.

Each color means a topic, and the x-axis shows a specific time, which is every month, and the y-axis shows the number of objects belonging to the topic. In other words, if the area of the topic becomes thicker at a specific point in time, it can be interpreted that the number of articles corresponding to the specific topic has increased at that point.

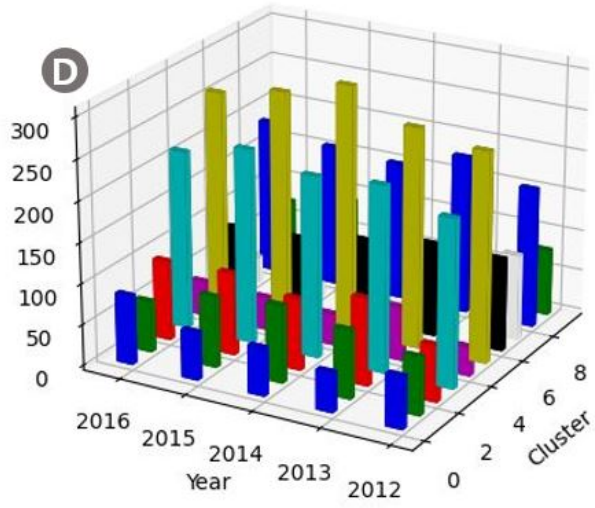


Figure 4.5: 3D Plot

In addition to the stack area plot displayed as the main, the UI also visualizes 3D Plot, as shown in Figure 4.5 Ⓔ on the additional screen. Unlike the Stack area plot, this 3d plot visualizes the topic modeling result every year, and it can be rotated 360 degrees so that this plot can be analyzed at various points of time or viewpoints that users want.

Since both 2D and 3D analysis are possible on our UI, users have the advantage of being able to select plots suitable for individual purposes. Use cases using this panel will be covered in detail in Chapter 5.

### 4.2.4 Topic Details


Figure 4.6: Topic Details

*Topic Details* Ⓔ in Figure 4.6 is a panel displayed simultaneously with *Topic Trend*

*Viewer* Ⓒ. As the name suggests, it contains the detailed information of each topic and shows up to 10, 15, and 20 topic information according to the desired topic number entered in *Parameter Panel* Ⓐ. Each list shows the number of articles belonging to a topic and the top 10 keywords representing the topic according to the computations explained in chapter 3. This enables users to check what essential keywords of each topic are, not explicit topics such as politics, society, and sports. Using this, users can intuitively know what articles each topic contains only by the keywords shown in Ⓔ when they choose the topic of interest in Ⓒ.

### 4.2.5 Article Viewer

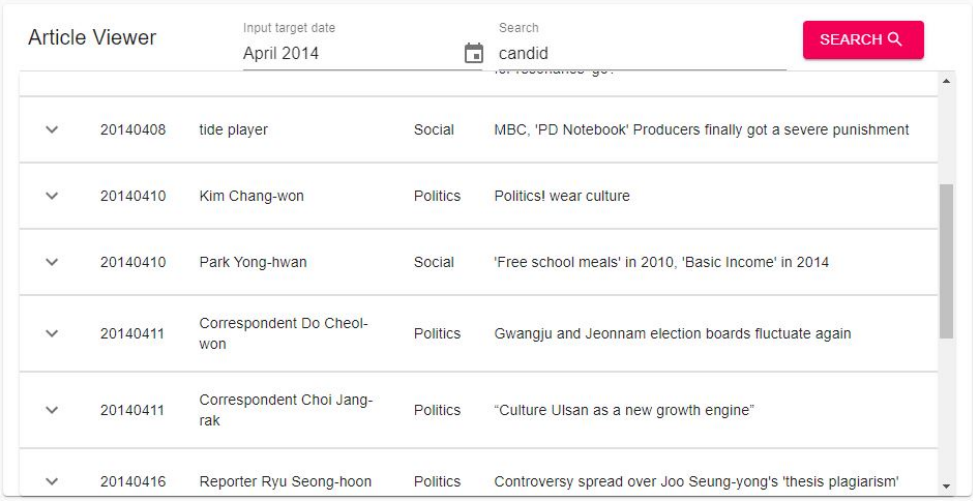


Figure 4.7: Article Viewer

*Article Viewer* Ⓕ is the last panel of the interface. After the users checked the change of article numbers in each topic over time and the top-$N$-keyword of each topic through the previous panels, the user can know the full text and details of the article directly from this panel. As in Ⓐ and Ⓑ, time and word input can be entered as a parameter.

As shown in Figure 4.7, the list of articles is sorted in order, and the time, author, main category, and title are displayed. The user can click the arrow displayed on the left to make the whole text visible, as displayed in Figure 4.8, and the details are checked through it,

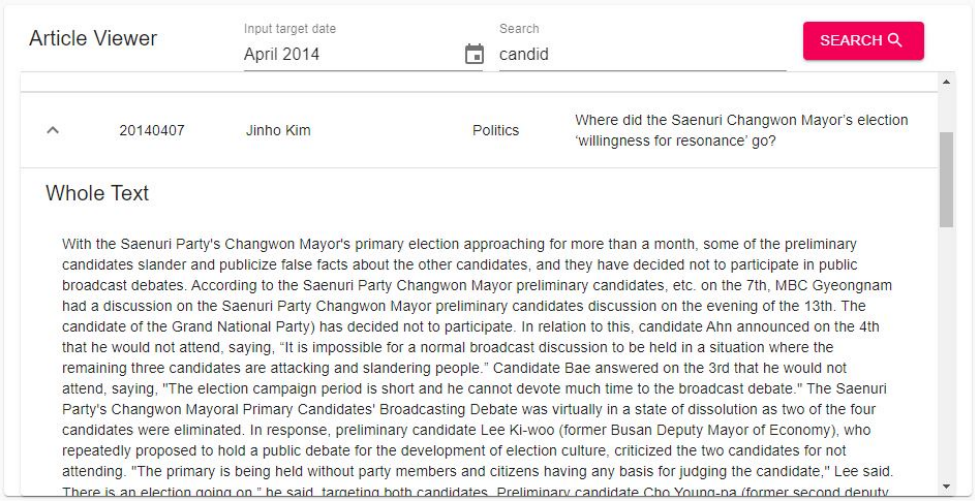


Figure 4.8: Whole artilce view from Article Viewer

## 4.3 Implementation Details

The visual interface is implemented in Javascript, CSS, and HTML using the framework React.js for front-end and rendering modules. The back-end computational modules consist of Python. API calls on a Flask application conduct all data exchanges between the server and the interface. The input data is pre-processed in a separate Python program and stored in a database.

# CHAPTER 5
# EVALUATION

This chapter illustrates the quantitative and qualitative evaluation results. We focus on the following performance criteria: (a) Does the topic modeling results look proper? (b) Is the processing speed of the method applied in the back-end fast enough to be used in the interactive interface? (c) In what situations can this study be used?

## 5.1 Data Sets

We use the Korean news dataset, which can be downloaded from the National Institute of Korean Language. All the data are written in Korean and mainly cover articles that occurred in Korea.

The initial dataset consists of about 3.5 million articles from 2009 to 2018. The primary categories are labeled in eight topics: Politics, Economy, Society, Life, IT/Science, Entertainment, Sports Culture, and Beauty/Health.

For the efficiency of the experiment and the visual interface development, about 100 articles from 2010 to 2018 are randomly selected for each month, and finally, about 10,800 sample article dataset is constructed.

In addition, many NLP models are based on English, we translated the sample dataset using Google Translate API.

Finally, pre-processing was also performed to obtain relationship information between articles and words of this sample dataset to make a TF-IDF matrix using a bag-of-word model.

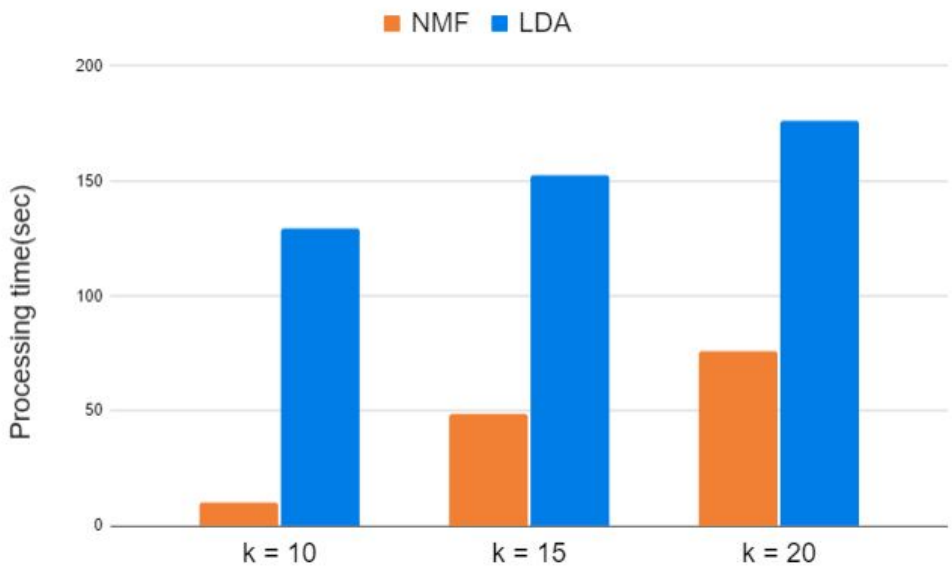


Figure 5.1: Processing time comparison with LDA. Desired lower rank is specified as $k$.

## 5.2 Quantitative Evaluation

### 5.2.1 Evaluation Setup

First, we set LDA, one of the most popular topic modeling methods, as our baseline model. Although accuracy and processing speed should be evaluated to compare model performance, past studies show that NMF and LDA outperform other methods under the same conditions as our visual interface [27] [26].

We perform all experiments for quantitative evaluations at least ten times each on a machine equipped with two Intel(R) Xeon(R) CPU E5-2680 v3 CPUs 2.5GHz and 378GB memory.

### 5.2.2 Running time

As mentioned at the beginning of the chapter, processing speed is essential in an interactive interface. No matter how accurate the processing result is, it is not suitable for interactive tools if the processing speed is slow. Users will not prefer an interface that requires more than a minute processing speed each time a parameter is changed. In the case of LDA, our

comparison method, a difference in the average topic coherence performance from NMF is not severe when the number of topics is small at ˜20, as mentioned above. However, as shown in Figure 5.1, the results of the topic modeling experiment for the three types of topic numbers 10, 15, and 20 provided by the interface show that the average time is 12 times longer at $k = 10$, 3.14 times longer at $k = 15$, and 2.3 times longer at $k = 20$.

In particular, Table 5.1 shows that LDA takes at least 2 minutes regardless of the value of $k$ and is not appropriate for use in an interactive interface.

Table 5.1: Processing time comparison result of two methods.

| | $k = 10$ | $k = 15$ | $k = 20$ |
|---|---|---|---|
| LDA | $125.59_{sec}$ | $152_{sec}$ | $176_{sec}$ |
| NMF | $10.4_{sec}$ | $48.53_{sec}$ | $76_{sec}$ |

## 5.3 Qualitative Evaluation

We qualitatively demonstrate our research's practicality by following actual use cases. These examples suggest where and how this interface can be used.

### 5.3.1 Use Case Study

**CASE 1 - Identify critical events** The first possible use case is identifying important keywords or events for a certain period of time using our visual interface. We can track past events using news data because it records and maintains events daily. However, it is not easy to identify what major events have occurred among large amounts of data.

There is a method of using external media, but there is a disadvantage in that reliability and objectivity are low. However, if we use the interface, we can see at a glance what significant events have occurred in a short period of one or two years and what important keywords have occurred through the event.

To demonstrate this use case, we conducted topic modeling to check what major events occurred in Korea during 2014 and obtained the plot shown in Figure 5.2.

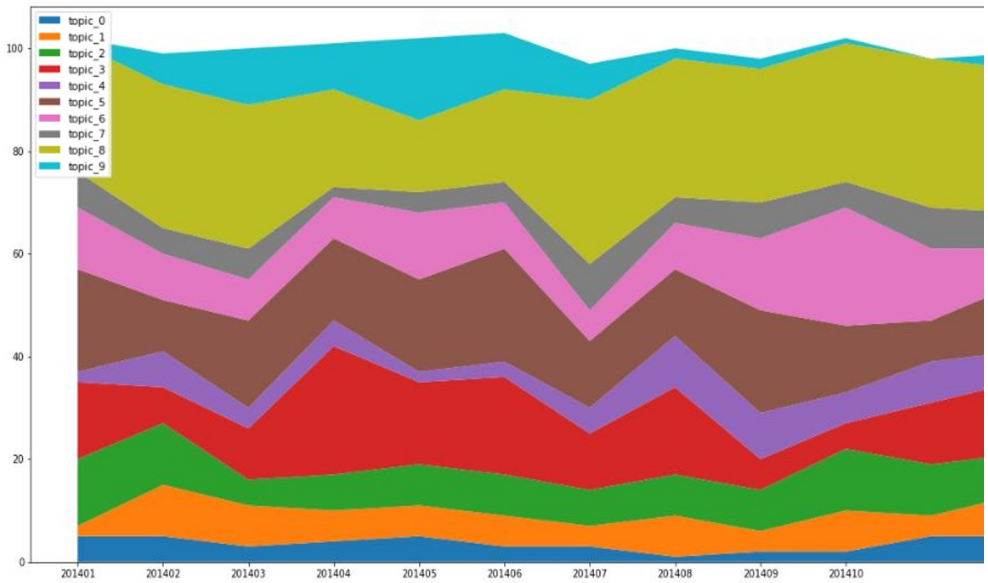


Figure 5.2: The stack area plot of Case study 1 : Topic 3 and topic 9 are displaying the increasing trend in April 2014 and May 2014

Figure 5.2 shows that topic 3's width increased rapidly in April 2014, and we can check the topic keyword lists, ['sewol', 'accid', 'ferri', 'polic', 'disast', 'family', 'ship', 'rescu'], are shown as keywords of that topic. This indicates a terrible incident in April 2014. A ferry carrying high school students on a school trip sank, and this incident killed 304 people. In 2014, this incident was handled with considerable interest in Korea.

In addition, topic 9 is also increasing in May 2014. When we check the keywords of the topic, it displays ['parti', 'elect', 'governor', 'local', 'polit', 'mayor', and 'vote']. We can say this topic indicates an event related to the election through this list. In fact, there were simultaneous nationwide local elections in Korea from the end of May to the beginning of June 2014.

As such, major events can be easily identified through topic modeling and the plot without reading all the lists and texts of articles.

**CASE 2 - Identify Sub-category** Contrary to CASE 1, if the time range is set wider and the initial category range is set narrower, the users can identify which subtopics were importantly handled within one category for a given time range. Moreover, users can classify the datasets or articles in detail. They can explain how the influence of each topic changes over time by using supported visualization methods such as 3d plots and stack area plots.

This finding detailed topics will also take a lot of time and cost and lose accuracy without using our UI.

Among the datasets from 2010 to 2018, we conducted a topic modeling experiment on the dataset corresponding to Sports. As a result, we identified several representative topics as follows :

1) Golf-related subtopics represented by [‘golf’, ‘tour’, ‘genesi’, ‘cours’, ‘swing’, ‘money’, ‘prize’, ‘wood’, ‘million’, ‘golfer’]

2) Pyeongchang Winter Olympics-related subtopics represented by [‘olymp’, ‘pyeongchang’, ‘winter’, ‘host’, ‘gangwondo’, ‘ioc’, ‘committe’, ‘ice’, ‘govern’, ‘cultur’ ]

3) Korean major league players-related subtopics represented by [‘ryu’, ‘hyunjin’, ‘dodger’, ‘yoon’, ‘seokmin’, ‘pitcher’, ‘pitch’, ‘era’]

We can confirm that subtopics clearly distinguished by these results, we are also able to analyze what sub-themes are important in the sports category during the given period. When news platforms subdivide and classify categories, they can classify by reflecting the interests of users by using this sub-topic information.

**CASE 3 - Understand relationships between topics** Finally, using this UI, we can check the relationship between topics. More specifically, we can see how critical events in one topic affect the other at a critical time point.

In real life, where we live, one event often has a ripple effect that causes another. For example, events such as the Olympics and the World Cup in sports bring about the nation’s economic growth. However, it is challenging to analyze the relationship because this is the

domain of predictions beyond the scope of statistic analysis.

However, our novel UI allows us to achieve some of these objectives. This is because news data includes all records from the whole topic daily. In other words, in the case of events that have occurred in the past or have similar cases, it is possible to predict what direct and indirect effects may occur by analyzing past event's news articles.

Let us assume that we want to know the effects of the 'election' on our lives. If we try to investigate and analyze using the whole data set, it takes much time, and objectivity is also low, making it a very inefficient task. However, since the election is a periodic event, our UI can perform the same task very efficiently if we focus on past election-related articles.

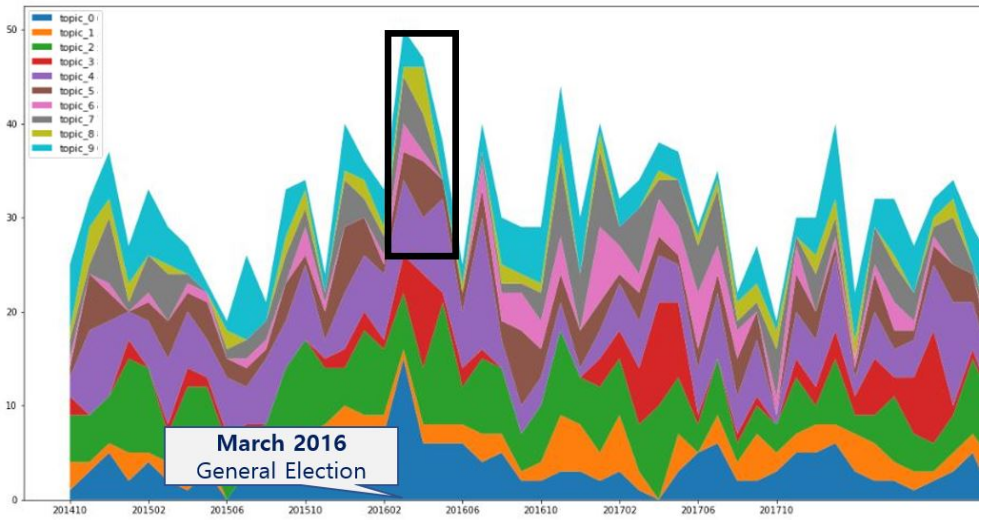


Figure 5.3: The stack area plot of Case study 3 : Topic 4,5,6, and 8 are increasing after the general election in March 2016.

First, the initial parameter value is set to use the interface. In order to search for articles related to the recent general election, which is an election for the members of the National Assembly in Korea, the range is set from October 2014 to September 2018, and we input the keyword 'election ' to get the articles related to the election.

When we look at the plot of March 2016, when the general election was held in Korea, we can see that the number of articles on topic 0, which is represented by keywords related to the Republican Party in Korea ['rep', 'saenuri', 'assembl', 'parti', 'opposit'], and topic

3, which is represented by keywords related to the Democratic Party of Korea, increases.

Not only this, but through changes in the graph from April, when the election was over, we can also observe how the election has direct or indirect effects on our lives. Figure 5.3 displays the rapid changes in topic 4, topic 5, topic 6, and topic 8 after the election.

The top keywords of each topic are :

(1) Topic 4 : [‘citi’, ‘gwangju’, ‘budget’, ‘local’, ‘construct’, ‘citizen’, ‘urban’],

(2) Topic 5 : [‘student’, ‘school’, ‘univers’, ‘educ’, ‘admiss’, ‘subject’, ‘grade’],

(3) Topic 6 : [‘nuclear’, ‘north’, ‘trump’, ‘power’, ‘plant’, ‘missil’, ‘korea’, ‘korean’]

(4) Topic 8 : [‘car’, ‘vehicl’, ‘motor’, ‘hybrid’, ‘electr’, ‘automobil’, ‘technolog’].

The keywords belonging to Topic 4 are about the development of regions related to the party that won the election, Topic 5 is about education, Topic 6 is about the relationship with North Korea, and Topic 8 is about the eco-friendly automobile industry. Through this, we can demonstrate that the results of the election lead to unequal development and interest in certain regions in Korea, affect the education system and the situation in North Korea, and finally, increase interest in specific industries such as the eco-friendly automobile industry.

It is possible to predict how a particular event affects other areas with this usage. It is an analysis above the level of statistics, and it is also the main contribution of this study. It can be practically used in various situations such as stock investment, policymaking, and business item selection.

# CHAPTER 6
# CONCLUSION

As news data's scale and complexity increase rapidly, research for topic modeling to discover semantic relationships using news data and research to visualize news data have been continuously conducted. However, in this thesis, we conduct research that includes all three concepts : (1) News data, (2) Topic modeling, and (3) Visualization, and present a prototype visual interface that can interactively utilize the result of the topic modeling and visualization.

Also, We demonstrate the NMF method's superior performance and apply it as a back-end algorithm for an interactive interface. The results of topic modeling and changes over time of topic are effectively visualized through stack area plot, 3d plot, and top N keyword.

Moreover, the study presented examples that could be used in practice rather than just in theory through the use cases.

## REFERENCES


[1] Z. Tang, X. Zhang, and S. Zhang, "Robust perceptual image hashing based on ring partition and nmf," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 3, pp. 711–724, 2013.

[2] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," *Genome research*, vol. 13, no. 7, pp. 1706–1718, 2003.

[3] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[4] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating non-negativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

[5] T.-J. Kim, "Covid-19 news analysis using news big data: Focusing on topic modeling analysis," *The Journal of the Korea Contents Association*, vol. 20, no. 5, pp. 457–466, 2020.

[6] Q. Liu *et al.*, "Health communication through news media during the early stage of the covid-19 outbreak in china: Digital topic modeling approach," *Journal of medical Internet research*, vol. 22, no. 4, e19118, 2020.

[7] T. Rajasundari, P. Subathra, and P. Kumar, "Performance analysis of topic modeling algorithms for news articles," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, pp. 175–183, 2017.

[8] K. Svensson and J. Blad, *Exploring nmf and lda topic models of swedish news articles*, 2020.

[9] D. Cohn and T. Hofmann, "The missing link-a probabilistic model of document content and hypertext connectivity," *Advances in neural information processing systems*, vol. 13, 2000.

[10] T. Griffiths and M. Steyvers, "Prediction and semantic association," *Advances in neural information processing systems*, vol. 15, 2002.

[11] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[13] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond svd," in *2012 IEEE 53rd annual symposium on foundations of computer science*, IEEE, 2012, pp. 1–10.

[14] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1, pp. 177–196, 2001.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] M. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, 2011.

[17] Y. Zhang, M. Chen, D. Huang, D. Wu, and Y. Li, "Idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization," *Future Generation Computer Systems*, vol. 66, pp. 30–35, 2017.

[18] J. Eisenstein, B. O'Connor, N. A. Smith, and E. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 1277–1287.

[19] S. Sizov, "Geofolk: Latent spatial semantics in web 2.0 social media," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 281–290.

[20] K. Choi, J. H. Lee, C. Willis, and J. S. Downie, "Topic modeling users' interpretations of songs to inform subject access in music digital libraries," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2015, pp. 183–186.

[21] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[22] D. Greene and J. P. Cross, "Exploring the political agenda of the european parliament using a dynamic topic modeling approach," *Political Analysis*, vol. 25, no. 1, pp. 77–94, 2017.

[23] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[24] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, pp. 1–22, 2016.

[25] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized, parts-based representation," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, IEEE, vol. 1, 2001, pp. I–I.

[26] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[27] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 952–961.

[28] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE, vol. 1, 2012, pp. 174–180.

[29] Y. Suh, J. Yu, J. Mo, L. Song, and C. Kim, "A comparison of oversampling methods on imbalanced topic classification of korean news articles," *Journal of Cognitive Science*, vol. 18, no. 4, pp. 391–437, 2017.

[30] A. Balahur and R. Steinberger, "Rethinking sentiment analysis in the news: From theory to practice and back," *Proceeding of WOMSA*, vol. 9, pp. 1–12, 2009.

[31] A. I. Bento, T. Nguyen, C. Wing, F. Lozano-Rojas, Y.-Y. Ahn, and K. Simon, "Evidence from internet search data shows information-seeking responses to news of local covid-19 cases," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11 220–11 222, 2020.

[32] P. K. Narayan, "Oil price news and covid-19—is there any connection?" *Energy Research Letters*, vol. 1, no. 1, p. 13 176, 2020.

[33] N. L. Kolluri and D. Murthy, "Coverifi: A covid-19 news verification system," *Online Social Networks and Media*, vol. 22, p. 100 123, 2021.

[34] J. Choo *et al.*, "Visirr: Visual analytics for information retrieval and recommendation with large-scale document data," in *2014 IEEE conference on visual analytics science and technology (VAST)*, IEEE, 2014, pp. 243–244.

[35] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 1992–2001, 2013.

[36] H. Kim, B. Drake, A. Endert, and H. Park, "Architext: Interactive hierarchical topic modeling," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 9, pp. 3644–3655, 2020.

[37] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: Visualizing theme changes over time," in *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, IEEE, 2000, pp. 115–123.

[38] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE transactions on visualization and computer graphics*, vol. 8, no. 1, pp. 9–20, 2002.

[39] J. A. Wise *et al.*, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proceedings of Visualization 1995 Conference*, IEEE, 1995, pp. 51–58.

[40] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.

[41] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, "Text mining using non-negative matrix factorizations," in *Proceedings of the 2004 SIAM international conference on data mining*, SIAM, 2004, pp. 452–456.

[42] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.