CS474 term project paper

Jihee Park
KAIST
Daejeon, Korea
j31d0@kaist.ac.kr

Junseop Ji
KAIST
Daejeon, Korea
gaon0403@kaist.ac.kr

Soyoung Yoon
KAIST
Daejeon, Korea
soyoungyoon@kaist.ac.kr

ABSTRACT

Due to the massive increase of news articles in the internet, the importance of topic analysis and issue tracking is growing. However, the massive amount of data makes people hard to do the work manually, so the automatic process held by the machine is needed. In this paper, we suggest an automatic news analysis process, which consists three steps: *1. trend analysis*, *2. on-issue event tracking*, and *3. off-issue event tracking*. For trend analysis, we use LDA with NER promotion, and for on-issue and off-issue tracking, we use DBSCAN and several libraries. After the experiment, we see that our trend analysis model clusters the news articles by topics very well, and event tracking models find out the events for each issue(topic). our code can be found on the github repository. [1]

KEYWORDS

topic modeling, event tracking, news analysis

1 INTRODUCTION

Online contents has grown big recently, and people are viewing more news on the internet. However, due to the heavy amount of data(news), it has a limitation to categorize the news manually. Also, it is almost impossible to track the events related to the issue while looking all the articles.

Looking at all yearly issues, we can see that the news articles can be clusted into some issues. For example, in the case of the year 2017, there was a lot of news related with the former president and her political crimes. The term *Trend Analysis* means clustering those kind of news articles and analyze the clusters. And, for each issue, we can see the *events* related to the issue and we can make the timeline of the main events for an issue. In detail, there are two kinds of event: the first one is the event which is directly related to an issue, and the second one is not directly linked, but topically related issue. In the paper, we call the first one as *"on-issue event"*, and the second one as *"off-issue event(related-issue event)"*

In this paper, we suggest a method to analyze the trends and track the events by three steps: 1. *Trend Analysis*, 2. *On-issue Event Tracking*, and 3. *Off-issue Event Tracking*.

After doing all the progress, we could analyze the yearly trends of Korea from 2015 to 2017. Also, we tried to track down some important issues extracted from above. We captured four quarterly events for each issue, and also tracked several off-issue events.

Our research is important and effective because we minimized the human(manual) efforts throughout the progress, for summarizing the 270K news articles. The methods we suggest in this paper can be widely used in yearly trend analysis, not just news, but marketing, research, or the other fields as well.

[1] https://github.com/soyoung97/Topic_modeling-Issue_Tracking

CS474 Term Project, 2019 Fall, KAIST

2 TREND ANALYSIS

2.1 Data Preprocessing

2.1.1 Data Format. As described in the READ.ME of data provided, The targeted data is from the Korean Herald, National Section news. The period of the dataset is from 2015 to 2017. The Crawled date of the dataset is 2018-10-26. Data format is Json, and there are total of 6 data headers - title, author, time, description, body, and section. Total of 23769 news are included in this dataset.

2.1.2 Load Data. In order to load the data, the instructions recommended at READ.ME are followed. Pandas library is used for better storing and access of the news text.

2.1.3 Libraries Used. For trend analysis, we used pandas, gensim, nltk, and neuroner python libraries. The install requirements are found on install.sh of the github repository.

2.2 Previous Approaches

Issue trend analysis can be seen as a part of Topic modeling. By searching fields of recent Topic modeling, LDA has shown to have good performance. As a result, LDA is used as a baseline algorithm for trend analysis. A recent study(2018) on Topic Modeling shows that Topic Quality improves when Named Entities are promoted.[3] This paper proposes 2 techniques: 1. Independent Named Entity Promoting and 2. Document Dependent Named Entity Promoting. Independent Named Entity Promoting promotes the importance of the named entities by applying scalar multiplication alpha to the importance of the named entity word. Document Dependent Named Entity Promoting promotes the importance of the named entities by setting the weights of the named entities as maximum term-frequency per document. For Independent Named Entity Promoting, the value of alpha can be changed flexibly, but results conducted by this paper shows that setting alpha as 10 showed the best results. We take advantage of this paper's idea on Independent Named Entity Promoting and implement Named Entity Promoted Topic Modeling by LDA.

2.3 Experiments

2.3.1 Data Tokenization. Several attempts were taken before we finalize the way Tokenization was done. Doing lemmatization was not always good. At first try, Lemmatization(converting words into base forms) and removal of stopwords were conducted before we run the LDA algorithm and extract Named Entities. We thought that converting words into base forms and reducing the total vocabulary size would increase the performance of topic modeling. Stopwords were taken from nltk.corpus.stopwords.words("english"), and lemmatization functon was taken from gensim.utils.lemmatize, and then res.append(lemmatize(raw_text, stopwords=stopwords)). But after we do lemmatization, remove stopwords, and tokenize the data, no Named Entities were extracted from the preprocessed corpus. We think the reason for this is as follows. First, words are all converted into lower case when we do lemmatization. This makes the Named Entitiy Recognition system(NER system) to work poorly because we have removed some of the original information(i.e. Upper case information), and word that starts with an upper case has a high probability that it is a "Proper pronoun", or "Unique word". We lose this sign of information. Second, words are transformed into their base forms, limiting NER system to detect specific words. There also could be cases that the words are transformed into meanings other then their original meanings. For example, "Cooking" and "Cooker" are both converted into "cook" when they are lemmatized, and this makes the word to lose the original information. Third, original relationships between words are lost, because of the removal of stopwords. When we do NER, we have to do the POS tagging of the sentence and then input both the word sequence and the POS sequence of the text. But when we artificially remove stopwords and then do NER, original relationships between words are disrupted and broken. This limits NER system to perform well.

For these 3 reasons, we decided to not apply lemmatization for tokenization, because lemmatization lose information about the original text. We decided to just use word_tokenize from nltk.tokenize, do POS tagging and then do NER.

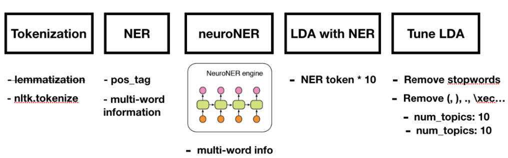

Figure 1: Overall flow of Trend Analysis process.

2.3.2 Extract NER. By using ne_chunk from nltk and pos_tag from nltk.tag, we extracted Named entities from the original news dataset. NER also extracts multi-word information of Named Entities other than just classifying whether a word is a named entity or not, so we decided to use that information. We store single-word Named Entities and multi-word Named Entities separately. As a result, NER and multi-word extraction of NER are both processed.

Below figure is the topic modeling result(of all time lengths from 2015 to 2017) WITH NER Promoting and WITHOUT NER Promoting. We can see the difference between those two results, and we can conclude topic modeling with NER promoting shows better performance.

2.3.3 Improvement - apply neuroNER instead of nltk's Named Entity Recognition. The topic modeling paper that we referenced used neuroNER for Named Entity Recognition. NeuronNER is an easy-to-use program for named entity recognition based on neural networks presented in emnlp 2017. [1] This neuroNER tool is trained on CONLL2003 dataset and recognizes four types of NE: person, location, organization and miscellaneous. NeuroNER also extracts multi-word information, so we use this multi-word information just as the previous NER did. Instead of using ne_chunk(pos_tag(preprocessed_text), binary=True), we change NER extraction to use below.

```
nn = neuromodel.NeuroNER(train_model=False,
     use_pretrained_model=True)
nn.predict(preprocessed_text)
```

to extract Named Entities from the text.

2.3.4 Run LDA with neuroNER promoting. First, we split the dataset each year. Then, get tokens for each document with promoted NER frequency (X 10). With this corpus, run the LdaModel with num_topics of 10 and num_words of 30 to 50. At first try, we directly ran LDA on NER boosted news dataset. but with this approach, we found out that stopwords are classified as top(important)words according to the result of LDA. So we decided to remove stopwords after all the preprocssing(including NER weight promoting)are done. The timing of removal of stopwords are important, as removing stopwords before NER will affect the NER result(Removal of stopwords before POS Tagging will affect the POS Tagging result). Stopword removing are done right before feeding the tokens into LDA. After the removal of stopwords, we could see that the results were much better.

2.3.5 Tuning LDA hyperparameters. We set num_topics to 10 for LDA becuase we need to extract top 10 important issues from each year. At first, we decided to train the LDA model with num_topics of 10 and num_words of 15. But the results were not very explainable. Also, the only removed word was the stopword after tokenization. Therefore non-ascii character, or unrelated words such as `" \xec ' ( ) . ,` were introduced in the topic result. To extract useful information, we removed those unuseful information and increased num_words for each topics to 50 to see more related words including each topic. First we set chunk_size to 2000, num_iterations to 4000, and alpha to 'auto'. We changed chunk_size to 4000 and increased num_iterations, and see if the result improved. But there was no significant change on the results. We finally decided to set num_topics to 10, chunk_size to 4000, iterations to 500, and passes to 30.

2.3.6 TroubleShooting. In order to increase the performance of Topic modeling, various approaches were taken. The first trial was to divide news dataset into given sections then do LDA modeling for each year, for each topic. But this approach can not detect the top 10 trending issues. Increasing the total topic size to more than 10 makes us difficult to analyze which topic is the top 10 most trending topic. This happened to be the same problem when we increase the total topic_size to more than 10. But, setting the total topic_size to 10 also has problems. By setting the total topic_size to 10 for each year, many topics can be concatenated into one. For this case, we filter out the majority topic by looking at the extracted tokens for each topic. Also, for the herald dataset, words related for "Korea" (Korean, North Korea, South Korea, Korean, ...) are used very frequently accross all topics, so the Topic analysis result for this also showed great frequency of words related to "Korea", making "Korea" unuseful for topic detection. Also there were topic clusters that were hard to analyze the keyword. Also, NER was good at extracting **multi-words**, but was not good at extracting **triple or more words**. For example, the neuroNER output of "Moon Jae-in eat food" was "Moon Jae", not 'Moon Jae-in". We could see the

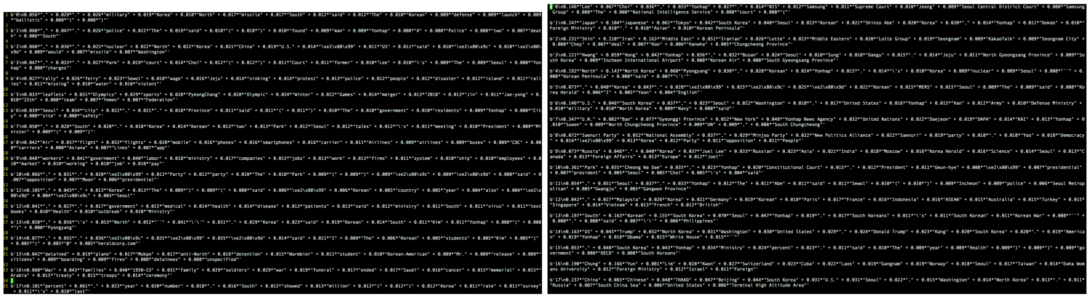

Figure 2: Topic modeling result before/after NER

inherent limitations when we try to extract 2 or more words just by using NER. To overcome those problems, more data preprocessingand multi-word extraction should be done. Also, If we link the corresponding news article that best represent a particular topic, it will make Topic modeling result more analyzable. Also, one inherent limitations with LDA topic modeling is that the result is given as a set of words. It is hard to analyze and manually label the overall topic just by looking at the set of words. If we can train a bigram, or n-gram language model with the words at each topic and generate topics out of the model, it will be much better analyzable.

3 ON-ISSUE TRACKING

For on-issue tracking, we first divide news articles monthly. Then we classify news articles in each month group into 10 issue categories. For each classified group, each article's 5W1H(when, where, who, what, why, how) is extracted and counted. The frequencies are used to extract the most relevant news title for each month.

Figure 3 shows the structure of the on-issue tracking process.

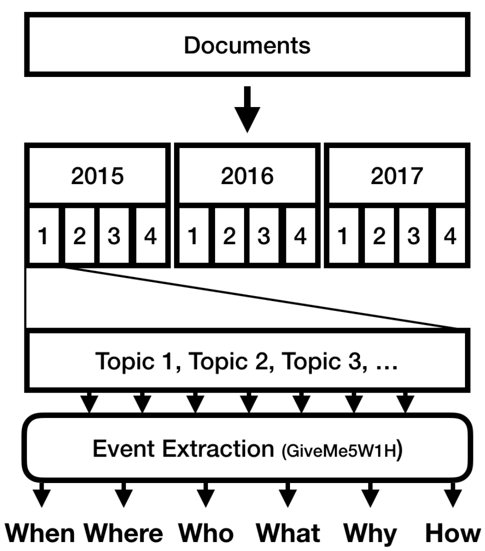

Figure 3: A brief diagram of on-issue tracking process.

3.1 Monthly Division

We divided all news articles monthly. The groups contain news articles those are written in *January 2015, Febuary 2015, . . . , December 2017, January 2018. 2018 Jan.* group contains only a few articles, so we decided to ignore the last group. The reason why we divided the data monthly is, the month is one of the standard in the field of yearly statistics analysis. For example, the issue about MERS started from May 2015, and ended in January 2016. If we divide yearly, there will be only one or two groups for extracting events. Else if we divide quarterly, there will be three or four events. We can extract eight or nine events from the period if we divide the events monthly, so this is just fit to make a reasonable result.

3.2 Articles in the Months Categorization

With LDA model we have trained at trend analysis project, we classify the documents in the month groups. If we give a tokenized sentence to the LDA model, the model outputs the probability for each group. We choose the group with maximum value, and assign the document to the group. So, for each month, there are 21 classified groups of news articles.

3.3 Event Extraction

For each group we divided from above, we extract the events with the approach of word frequency. For this step, we use a Python library called "giveme5W1H"[2]. The library is the state-of-the-art tool for extracting *when/where/who/what/why/how* features from the document. The library uses Stanford's CoreNLP library as its basic structure, and give analysis results when we give a title, lead, text, and a published date. We decided to use columns *title*, *description*, *body*, and a *time* from the given dataset as an input to get a result. For each group, we count the frequencies of each feature of the articles, and select the most frequent terms for each feature, treat them as a score. Then we extract a most relevant article from the monthly group; For example, if the term 'president' occurs twelve times and 'government' occurs six times as *Who* feature, the news article contains the term 'president' as *Who* takes double scores than the article about 'government'. The maximum score article's headline is assumed that it is representing the main event of the month.

We choose two yearly issues from the list, and do event extraction for each issue. For each month's result, we identify an event based on the result and align them on the timeline.

Table 1 is an example of on-issue tracking of the issue MERS.

Month	Event(headline)
2015.06	S. Korea confirms 3 more MERS cases, total rises to 18
2015.07	S. Korea reports no new MERS cases for 17th day
2015.08	Park gives appointment letter to new health minister
2015.09	Moon stakes leadership on party reform
2015.10	61 isolated after last MERS patient rediagnosed

Table 1: The example of on-issue tracking of the issue MERS.

4 OFF-ISSUE TRACKING

For off-issue tracking, we first categorize topics given as Trend analysis part. In this section, we denote a document as sequence of tokens plus its created time $\mathbb{D} := (\Sigma^+, t)$, when $t \in \mathbb{R}$ (timestamp of creation time). and the set of document of topic a as $\mathbb{T}_a \in \mathcal{P}(\mathbb{D})$.

4.1 BoW Extraction

In first, we have to extract document in some space which we can analyze quantatively. We use BoW as morphism from document space to vector space $\mathbb{R}^N$, which we can analyze similarity of document. In addition, we add one more dimension to give information of document creation time. From pre-calculated set of tokens $\Sigma := \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$, our transformation $b : \mathbb{D} \to \mathbb{R}^{n+1}$ is defined inductively as

$$\begin{cases} b([], t) := t * e_{n+1} \\ b(\sigma_i :: tl, t) := e_i + b(tl, t) \end{cases}$$

Then, morphism from $\mathbb{T}_a \in \mathcal{P}(\mathbb{D})$ to $\mathcal{P}(\mathbb{R}^{n+1})$ is naturally induced from b as $\phi(\mathbb{T}_a) = \{b(d) | d \in \mathbb{T}_a\}$

4.2 Relation Between Semantic of Document and BoW

We know that there are documents and events which have similar meaning, but we cannot formalize it because we currently do not have model of language interpretation in metric space. But we can assume *such* space exists, i.e. there is an isomorphism $\phi : \mathbb{D} \to \mathbb{D}^\#$, when $(D^\#, d^\#)$ is metric space. It is not hard to assume this structure, since similar concept is already introduced as Entity comparison/Behavior comparison operator of Semantic algebra [4].

Our desired result is that b with euclidean distance successfully models $(D^\#, d^\#)$, but we cannot show it because we do not have constructive definition of $D^\#$. But if it has sufficient approximation, (bounded approximation) We can derive more interesting properties (such as bounded error from BoW to Event space, etc).

Definition 4.1. b has approximation of ϕ with bound K, ϵ iff there exists an Lipshitz continuous π with K that $d^\#(\pi(b(d)), \phi(d)) \leq \epsilon$.

4.3 Relation Between Semantic of Event and BoW

Once semantic of document is defined, we can build similar notion of event as metric space. To build such space, we first understand about relation between document and event.

- similar document refer similar event.
- similar event (even same event) may be refered by documents with far distance, but it is not arbitrarly far.

we can formulize this as logical formlua, with definition of $e : D^\# \to E^\#$. $((E^\#, e^\#)$ is metric space for event)

- if $d^\#(d_1, d_2)$ is sufficiently small, then $e^\#(e(d_1), e(d_2))$ is sufficiently small.
- when $e^\#(e(d_1), e(d_2))$ is small, it doesn't mean $d^\#(d_1, d_2)$ is small but is bounded.

begin with this fact, we can find very interesting property which generalize this: continuity.

Definition 4.2. e is Lipschitz continuous with K if and only if $e^\#(e(d_1), e(d_2)) \leq K d^\#(d_1, d_2)$.

We can check that if e is Lipschitz continuous with K_e, then above two property is satisfied. Also, it derives important fact: If we have an approximation of semantics with bounded error, then there also exists approximation of event with bounded error.

Theorem 4.3. *b has approximation of ϕ with bound K, ϵ, then there exists $\pi_e : \mathbb{R}^{n+1} \to E^\#$ s.t. $e^\#(\pi_e(b(d)), e(\phi(d))) \leq K_e \cdot \epsilon$. (it means b has approximation of $e \cdot \phi$ with bound $K, K_e \cdot \epsilon$)*

Although proof is directly derived from Lipschitz continuity, it emphasizes that if we have bounded approximation of document, then it guarantees bounded approximation of event.

4.4 Event Clustering

In this assumption about semantic of document ans event, we can build event clustering method. Before using techniques in R^{n+1}, we focus on how this clustering in R^{n+1} effects in $E^\#$.

Theorem 4.4. *if b has approximation of $e \cdot \phi$ with bound K, ϵ, then $e^\#(e \cdot \phi(d_1), e \cdot \phi(d_2)) \leq 2 \cdot \epsilon + K||b(d_1) - b(d_2)||$.*

Proof.

$$\begin{aligned} & e^\#(e \cdot \phi(d_1), e \cdot \phi(d_2)) \leq e^\#(e \cdot \phi(d_1), \pi_e(b(d_1))) + \\ & e^\#(\pi_e(b(d_1)), \pi_e(b(d_2))) + e^\#(\pi_e(b(d_2)), e \cdot \phi(d_2)) \leq \\ & \epsilon + e^\#(\pi_e(b(d_1)), \pi_e(b(d_2))) + \epsilon \leq \\ & 2 \cdot \epsilon + K||b(d_1) - b(d_2)||. \end{aligned}$$

□

It shows that, if we make good Vector transformation b, then it automatically guarantees bounded error for distance of extracted event, without construction of π, ϕ, e or any other. Begin with this fact, we derive constructive definition of partition for documents using approximated transformation b. To do that, we first define similarity relation for two documents.

Definition 4.5 (Similarity Relation). $\approx_{\mathbb{R}^{n+1}, \delta} \in \mathcal{P}(\mathbb{D} \times \mathbb{D})$ is defined as

$$d_1 \approx_{\mathbb{R}^{n+1}, \delta} d_2 \iff ||b(d_1) - b(d_2)|| \leq \delta.$$

Similarly, $\approx_{E^\#, \delta} \in \mathcal{P}(\mathbb{D} \times \mathbb{D})$ is defined as

$$d_1 \approx_{E^\#, \delta} d_2 \iff e^\#(e \cdot \phi(d_1), e \cdot \phi(d_2)) \leq \delta.$$

then $\approx_{\mathbb{R}^{n+1}, \delta} \subseteq \approx_{E^\#, 2 \cdot \epsilon + K \cdot \delta}$ holds by above theorem. Thus it is quite reasonable to use $\approx_{\mathbb{R}^{n+1}, \delta}$ to cluster events, instead of uncomputable relation $\approx_{E^\#, 2 \cdot \epsilon + K \cdot \delta}$.

Definition 4.6 (Transitive Closure). $\approx^*_{\mathbb{R}^{n+1}, \delta}$ is smallest relation on $\mathbb{D}$ that contains $\approx_{\mathbb{R}^{n+1}, \delta}$ and is transitive.

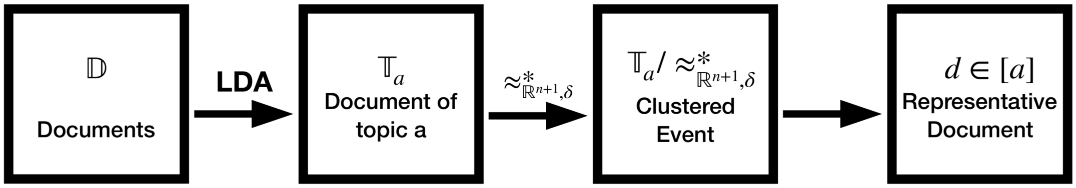

Figure 4: Overview of off-issue tracking process.

Then $\approx^*_{\mathbb{R}^{n+1},\delta}$ is reflexive, symmetric and transitive, which can be considered as equivalence relation. Then, we can partition documents with this equivalence relation.

Definition 4.7 (Partiton of $\mathbb{D}$). when $\approx$ is equivalence relation, $\mathbb{D}/\approx := \{[a] | a \in \mathbb{D}\}$, when $[a] := \{b \in \mathbb{D} | a \approx b\}$.

By substitute $\mathbb{D}$ to $\mathbb{T}_a$, finally we have $\mathbb{T}_a/\approx^*_{\mathbb{R}^{n+1},\delta}$ as successful approximation of event partition of topic a. Now, we are going to explain how most relevant description of event is extracted from each partition.

4.5 Extracting Representative Description

Now we have cluster of events (documents which describing events) $\mathbb{T}_a/\approx^*_{\mathbb{R}^{n+1},\delta}$, but we should return summary of events, because whole collection of documents are quite long to read and might have unnecessary information. So we have to extract *representative description* of tht event cluster. To extract target information from a document is well studied in information extraction field, and there are several method such as template-based information extraction, neural methods, etc. But in the case of several documents, it is hard to converge summary to cover all document's information, because existing works is not based on language semantic-based, so it is hard to generate summary statement between description of similar/same meaning.

For example, if one document describes the event happens "one day after of 12/7", and there are another document describe the event was happened "one day before of 12/9". Obviously, both description refer same day, but token-based approach (or pattern-based approach such as signal words) cannot handle this issue. Even with this disadvantage, above method is widely used because of its high performance (and due to challenges of semantic based information extraction method).

So, we decided to use event extractor for one document, but we design to choose representative document appropriately.

Definition 4.8 (Representative docuemnt). document $d \in [a]$ is *representative document* of $[a]$ when $\sum_{d' \in [a]} ||b(d) - b(d')||$ $\leq \sum_{d' \in [a]} ||b(x) - b(d')||$ for any $x \in [a]$.

It means that we choose to extract event from a document which has minimum difference between all other documents. After choosing representative document, we use Giveme5W1H framework[2] to extract description of event.

4.6 Implementation

To implement BoW transformation and document clustering, we use pandas and gensim for python. to calculate transitive closure and finding partition, we use DBSCAN algorithm. Parameters are adjusted by experiments on small set of documents. After that, extracting event description is done by Giveme5W1H framework.

5 EVALUATION

5.1 Trend Analysis Evaluation

To evaluate our result, we first try to show that our trend analysis works well. To do that, we collect other document with topic label. With these test set, its trend analysis result indirectly shows our accuracy of trend analysis.

5.2 Selecting Test Set

We use reuters data set. It consists of more than 9000 documents with more than 70 topics. But, the similarity of document is important because evaluation on very diffrent set of documents doesn't imply any meaningful result. To resolve that, we decided to extract 10 topics with most similarity between our dataset. It is achieved by calculating document similarity between reuters dataset and our news dataset.

To pick most similar topic, we compute maximum similarity within documents in topic and minimum similarity between reuters and news dataset. Due to largeness of dataset, we choose only subset of dataset to calculate minimum/maximum similarity bewteen groups. Similarity of groups are represented as graph in figure 5. Distance of nearest 10 topics in reuters dataset is shown as table 2. Also, there are significant difference of distances between nearest group and others, as shown as table 3.

5.3 Reuters Evaluation

With reuters dataset with 10 pre-classified topics, we generate LDA model for reuters dataset and make 10 topics. And we classified the topic with given label. In result, we successfully classified 7 topics from LDA result. It shows that our LDA model seems to work correctly. Precise result is shown as table 4.

we think that this evaluation is meaningful becuase the evaluation was conducted on well-known external **labeled** dataset sources, which proves the objectivity of the evaluation. In other words, this evaluation was not just done by manual judgement. Also, among those reuters datasets, we only evaluated those that showed close similarity values(similar trends) with our Herald dataset, thus increasing the accuracy.

5.4 Off-issue Tracking

For off-issue tracking, we just evaluated dunn index of clusters for many epsilon values of DBSCAN algorithm. If dunn index is high,

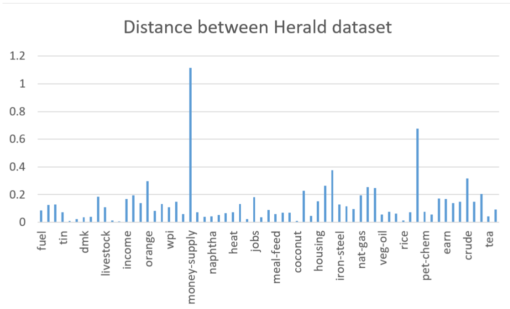

Figure 5: Distance between Herald dataset.

Topic	distance
yen	0.0061
lumber	0.0524
veg-oil	0.0566
strategic-metal	0.0576
carcass	0.0593
metal-feed	0.0606
gold	0.0622
ship	0.0681
cocoa	0.0718
oilseed	0.0731

Table 2: Distance of nearest 10 topics between Herald dataset.

Group	distance
nearest 10 topics	0.061
other topics	0.1964

Table 3: Difference between nearest topics and another.

Topic	matched label
0	Unknown
1	oilseed
2	veg-oil
3	gold
4	Unknown
5	Unknown
6	lumber
7	yen
8	oilseed
9	cocoa

Table 4: Matched topic and label.

then it means cluster has higher distance between cluster, and has lower distance in cluster. The result shows that smaller epsilon makes greater dunn index, but number of cluster is decreasing. So, we have to decide appropriate epsilon value for better result. Numerical result of calculation is shown as table 5.

Epsilon	max Δ_k	min $\delta(C_i, C_j)$	DI_m	# of cluster
3	8.122	192.0	24.26	5
5	22.23	163.2	7.34	12
7	52.72	138.4	2.625	16
9	106.7	102.1	0.9564	28

Table 5: Dunn index and number of cluster for internal evaluation.

6 CONCLUSION

Through Trend analysis, On-issue tracking and off-issue tracking, we were able to really apply on several nlp tools, and learn new features. As a result, we were able to successfully extract useful information based on Herald dataset. There were new ways, as well as limitations to our approach. Detailed output and analysis is listed on the appendix.

7 CONTRIBUTION

For this project, Soyoung Yoon took account of top 10 trend analysis and evaluation by the reuters dataset. She did experiments, coding, ppt-making and writing of the trend analysis and evaluation. Jihee Park took account of off-issue tracking. He did experiments, coding, ppt-making and writing of off-issue tracking and evaluation. Also, he organized the latex report template. Junseop Ji took account of on-issue tracking. He did experiments, coding, ppt-making and writing of on-issue tracking. All the ideas to solve each problems were initially proposed by each one that is taking part of it and throughly discussed by all other teammates. All teammates gave recommendations for improvements and helped each other for their part when it was having problems.

REFERENCES

[1] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)* (2017).
[2] Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5W1H: A Universal System for Extracting Main Events from News Articles. In *Proceedings of the 13th ACM Conference on Recommender Systems, 7th International Workshop on News Recommendation and Analytics (INRA 2019)*.
[3] Katsiaryna Krasnashchok and Salim Jouili. 2018. Improving Topic Quality by Promoting Named Entities in Topic Modeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 247–253. https://doi.org/10.18653/v1/P18-2040
[4] Yingxu Wang. 2013. On Semantic Algebra: A Denotational Mathematics for Natural Language Comprehension and Cognitive Computing. *Journal of Advanced Mathematics and Applications* 2, 2 (2013), 145–161.

A APPENDIX

A.1 Trend-analysis: 2015 top 10 trends

Non-informative words are excluded. Words that appear earlier in one topic represent higher probability than those which appear later.

(1) **Roh Moo-hyun's Oppression by NIS** Kim, Lee, Hwang, NIS, Seoul, Yonhap, Sung, Hong, Samsung Medical Center, Lee Wan, Lee Myung, said, POSCO, former, South Korea, Ministry of Health and Welfare, Kim Ki, Supreme Court, Shin, Seoul Central District Court, Yoon, The, Hwang Kyo, National Intelligence Service, Lee, Daejeon, Kim, National

Security Law, Kim Young, Kim Hyun, South Korean, Lee Byung, also, SMOE, He, Sung Woan, Kato, Kwon, Seoul National University Hospital, Ock Hyun, court, prosecution, Sung Wan, Seoul Central District ProsecutorsÓffice, Keangnam Enterprises, Chun Doo, I

(2) **Thaad placement** U.S., South Korea, Seoul, Washington, United States, North Korea, American, Iran, THAAD, Obama, South Korean, Yonhap, Barack Obama, Korea, said, China, Asia, Korean, “, Republic of Korea, White House, Asian, India, ś, Pentagon, State Department, Mark Lippert, Carter, Pacific, nuclear, The, Czech Republic, security, Lippert, also, defense, Iranian, Japan, Americans, ROK, USFK, U.S. Forces Korea, Han Min, Czech, Yun Byung, Congress

(3) **MERS** MERS, Seoul, South Korea, Yonhap, said, The, percent, Choi, South Korean, Cho, government, Gyeonggi Province, Seoul City, Lee Hyun, Health Ministry, Middle East, year, South Koreans, ś, people, Kang, also, Sewol, police, number, Claire Lee, Incheon, Saudi Arabia, country, million, KCTU, Middle East Respiratory, last, ´´, Seoul Metropolitan Government, ministry, public, first, Korean Air, Jeju, World Health Organization, Constitutional Court, Busan, city, Ock Hyun, disease, according, officials, cases

(4) **Saenuri Party and Park Geun-hye** Park, Park Geun, Saenuri Party, NPAD, National Assembly, New Politics Alliance for Democracy, Cheong Wa Dae, Moon, Ahn, party, Yoo, Park, said, Cabinet, The, Saenuri, Parkś, Cho Chung, Kim Moo, Assembly, opposition, government, Yonhap, Kim, Moon Jae, Chung, President, Rep. Moon Jae, Yeo Jun, ruling, -hye, would, also, presidential, Roh, Rep, Constitution, Kim Young, Jeong Hunny, Rep. Kim Moo, Wa Dae, public, Park Chung, political, lawmakers, Min Kyung

(5) **Korea compared with OECD** Korea, Korean, Koreans, Seoul, Joel Lee, The Korea Herald, English, The, said, OECD, French, Internet, I, British, Education Ministry, students, ASEAN, education, Asia, also, German, Europe, percent, Singapore, Canada, @, heraldcorp.com, government, year, India, Yoon Min-sik(minsikyoon@heraldcorp.com, Korea Herald, one, European, country, years, Yoon Min, Ock Hyun, France, would, It, people, EU, Polish, Greece

(6) **Statement of ICC** North Korea, North, North Korean, Pyongyang, Kim Jong, Kim, Yonhap, U.N., ś, “, said, United States, DPRK, nuclear, ´´, South Korea, U.S., U.N. Security Council, KCNA, Kim Il, Korean Central News Agency, North Koreans, The, Russia, SLBM, United Nations, leader, Hwang, Sung Kim, Security Council, U.N. General Assembly, Sony Pictures, International Criminal Court, Choe, also, Washington, Democratic Peopleś Republic of Korea, rights, Hwang Joon, missile, WorkersṔarty, human, talks, -un, State Department, Ban Ki, Workers Party, Ri, regime, country

(7) **Japan and sex slaves** Japan, Japanese, South Korea, Seoul, Tokyo, Korean, Abe, South Korean, World War II, Yonhap, Korea, Shinzo Abe, Park Geun, ś, Foreign Ministry, Ban, Yun, U.N., Dokdo, Koreans, “, UNESCO, Asian, Yun Byung, said, Park, ´´, Denmark, Danish, issue, women, Minister, New York, The, ASEAN, two, summit, also, Song Sang, Asia, Fumio Kishida, victims, talks, Northeast Asian, meeting, bilateral, wartime, history, countries, Ban Ki

(8) **Mount Geumgangsan-relationship with North Korea** North, Korean, North Korea, South Korean, Seoul, South, South Korea, North Korean, Koreas, Yonhap, Pyongyang, Korean War, South Koreans, Unification Ministry, ś, said, two, “, The, military, inter-, ´´, DMZ, Park Geun, Lim, Mount Kumgang, Song Sang, North Koreans, Kaesong, Korea, Cheonan, Yellow Sea, border, JCS, talks, Demilitarized Zone, Kim Jong, Kim Dae, Pakistan, government, North, The North, Mount Geumgangsan, Panmunjom, Red Cross, Northern Limit Line, official, Hong Yong, Gaeseong, Joint Chiefs of Staff

(9) **Indonesian Air Force** South Korea, South Korean, Yonhap, Seoul, Paris, Han, Navy, Turkey, Philippines, Indonesia, Syria, Air Force, Indonesian, U.N., said, Middle East, Malaysia, Defense Ministry, Turkish, Ebola, France, ś, Australia, Islamic State, The, DAPA, Peru, “, Foreign Ministry, South Koreans, Islamic, Han Min, Southeast Asian, Shin Hyon, Park, Vietnam, Thailand, Colombia, KAI, ´´, countries, Park Geun, Iraq, Lockheed Martin, Korean, Manila, Brazil, African, Saudi Arabia, Kuwait

(10) **South China Issue** China, Chinese, Beijing, South Korea, Russia, Russian, South Korean, South China Sea, Asia, Xi Jinping, AIIB, Seoul, Germany, Moscow, Xi, Yonhap, Europe, Taiwan, Park Geun, Hong Kong, Communist Party, Poland, German, FTA, World War II, Asian, Vietnam, Japan, ś, Western, Northeast Asia, Li, Li Keqiang, Mongolia, Cold War, KH, said, The, :, Ukraine, economic, Berlin, Eurasia, Asian Infrastructure Investment Bank, Soviet Union, Korea, Vladimir Putin, APEC

A.2 Trend-analysis: 2016 top 10 trends

(1) **Political Scandal about President Park** Park, Park Geun, Choi, Cheong Wa Dae, Seoul, Choi Soon, Yonhap, Cabinet, Park, Constitutional Court, Parkś, Kim, Sewol, said, president, President, South Korea, Seongju, Woo, National Assembly, The, presidential, -hye, Supreme Court, Wa Dae, Samsung, Chung, office, also, ś, Gwanghwamun Square, Woo Byung, Jung, Constitution, Lee, scandal, court, Blue House, Jeong, Jin, Choi Democratic Party of Korea, Seoul Central District ProsecutorsÓffice, former, Cho, Na

(2) **Zika Virus** Zika, Chinese, Yonhap, JCS, Incheon, South Korea, Ri, Jeju, Seoul, Vietnamese, Hong, Gyeonggi Province, Hong Kong, Celsius, said, KMA, GPS, MERS, Joint Chiefs of Staff, Coast Guard, KCDC, RFA, Korea Centers for Disease Control and Prevention, South Korean, The, Gender Ministry, Jejudo Island, Radio Free Asia, Jeolla, South Jeolla Province, Korea Meteorological Administration, Koreas, Suwon, Jeju Island, Northern Limit Line, Yang, Vietnam, Ministry of Public Safety and Security, West Sea, virus, North Gyeongsang Province, Punggye, Uzbekistan, DMZ, Demilitarized Zone, Jejudo, East Sea, Hangang River, Ministry of Science, ICT and Future Planning, Catholic

(3) **Seoul and Gyeonggi Province Issue** Korea, Korean, Seoul, Kim, Koreans, The Korea Herald, said, English, The, I, Ock Hyun, EU, –, Lee Hyun, students, @, heraldcorp.com, Gyeonggi

Province, Seoul Metropolitan Government, also, Seoul City, Education Ministry, Kim Da, Joel Lee, school, It, would, education, Justice Ministry, By, one, children, public, women, year, In, Oxy, years, government, Cho, people, Seoul National University, child, But, Gangnam, We

(4) **Relationship with Ban Kimoon and UN** Japan, Seoul, South Korea, South Korean, Korean, Japanese, Chinese, China, Yonhap, Tokyo, U.N., North Korean, Korea, Foreign Ministry, Ban, ś, South, said, Shin Hyon, Gaeseong, South Koreans, “, Park Geun, North Koreans, World War II, Shinzo Abe, Unification Ministry, UN, Ban Ki, ´´, The, Abe, Beijing, Wang Yi, government, Yun Byung, Dokdo, United Nations, two, Hiroshima, New York, Wang, Kaesong, US, also, U.S. Army, London, deal, last

(5) **Thaad and missile** North Korea, North, China, U.S., North Korean, South Korea, Pyongyang, Yonhap, Seoul, THAAD, Kim Jong, Washington, Beijing, ś, U.N., Kim, nuclear, U.N. Security Council, Russia, United States, said, “, South, Korean, ´´, South Korean, missile, Chinese, test, UNSC, sanctions, DPRK, North Koreans, The, Terminal High Altitude Area Defense, Security Council, Musudan, launch, also, Koreas, Korean Central News Agency, SLBM, Defense Ministry, military, State Department, Russian, North, KCNA, Korea, Yun

(6) **National Police Agency** France, Lim Jeong, French, Paris, African, police, Africa, The, Uganda, Ethiopia, Mongolia, Environment Ministry, Kuwait, Gyeonggi Province, Baek, British, Oxy Reckitt Benckiser, National Police Agency, said, Jeong, AI, victims, Kim Da, victim, Bak Se, Mongolian, humidifier, Seongnam, man, Seoul, FKI, PHMG, two, Air Koryo, Bucheon, Baek Nam, Baek, found, @, South Chungcheong Province, heraldcorp.com, Cefu, -yeo, kaylalim, A, ASEM, South Africa, death

(7) **South Korea-ASEAN Relationship** South Korea, Yonhap, South Korean, percent, Seoul, said, South Koreans, Busan, The, year, ASEAN, Philippines, ś, government, Claire Lee, “, Kang, number, million, country, also, ´´, Korea, people, last, Asia, Thailand, India, Southeast Asian, Vietnam, OECD, U.K., ministry, Singapore, showed, Pakistan, Middle East, workers, Malaysia, Internet, according, total, years, data, report, Cambodia, Gangwon Province, Organization for Economic Cooperation and Development, Incheon International Airport

(8) **Saenuri Party and general elections** Saenuri Party, National Assembly, Minjoo Party, Kim, Saenuri, Park Geun, Iran, party, Minjoo Party of Korea, People, Party, Moon, Ahn, Assembly, opposition, Chung, Lee, The Minjoo Party of Korea, Bae Hyun, The, Yeo Jun, Rep, said, Peopleś Party, Yoo, ruling, Seoul, Park, Moon Jae, Minjoo, Kim Chong, election, political, Iranian, Rep. Ahn Cheol, Justice Party, Tehran, parties, lawmakers, Yonhap, NIS, Democratic Party, Roh, parliamentary, Kim Moo, leader, former

(9) **US elections** U.S., South Korea, US, American, Seoul, United States, Trump, Obama, Yonhap, South Korean, Korea, Han, Navy, Korean, Washington, “, Army, said, Japan, Donald Trump, Republican, Air Force, Barack Obama, ´´, Han Min, Asia, Korean War, ś, USFK, America, military, Turkey, Pacific, Asian, Cuba, White House, Turkish, Americans, U.S. Forces Korea, Clinton, Congress, defense, The, Defense Ministry, Iraq, Republic of Korea, KAI, Nuri, DAPA, also

(10) **KATUSA** Korea, Lee, Korean, Hwang, Joel Lee, Europe, Hwang Kyo, Sri Lanka, Germany, German, European Union, Mexico, Seoul, European, Poland, Asia, New Zealand, Song, Britain, Polish, HIV, British, KATUSA, Korean War, House, DNA, Canada, The, Middle East, Gyeongju, AIDS, Sri Lankan, Office, Canadian, country, National Defense Commission, Netherlands, Colombia, Chun, Mexican, Greece, Ecuador, Minister, The Korea, Italy, Joel Lee / The Korea Herald, Jeon, Spanish, Gwangju, London

A.3 Trend-analysis: 2017 top 10 trends

(1) **PyeongChang Olympics** Japan, South Korea, Korean, Japanese, South Korean, Seoul, Tokyo, Yonhap, Kang, Korea, Russia, ś, Kang Kyung, Yun Byung, PyeongChang, Olympics, Busan, Germany, Shinzo Abe, Abe, World War II, “, Dokdo, said, East Sea, Ministry of Foreign Affairs, ´´, Yun, Russian, Foreign Ministry, Berlin, ministry, Koreans, Asian, Sri Lanka, The, Winter Olympics, Group of 20, government, also, Minister, Hamburg, German, Olympic, South Koreans, foreign, Sri Lankan, women, two, deal

(2) **Choi Sun-Sil gate and Presidential Impeachment** Park, Choi, Lee, Park Geun, Constitutional Court, Choi Soon, Yonhap, Seoul, Samsung, Park, Parkś, Samsung Group, Lee Jae, NIS, The, court, Seoul Central District Court, Chung, said, ś, Cho, President, former, president, Ock Hyun, impeachment, Kim Ki, Woo, -hye, scandal, Kim, presidential, Lee Kyu, Chung Yoo, Choi, Mir, team, Cho Yoon, National Assembly, trial, Samsung Electronics, Woo Byung, office, investigation, National Intelligence Service, South Korea

(3) **North Korea Relationship** North Korean, Kim Jong, North Korea, Kim, North, Pyongyang, South Korea, Yonhap, Seoul, Korean, South Korean, North Koreans, South, ś, said, Koreas, Kim Il, Korean War, Kaesong, leader, “, Korean Central News Agency, ´´, -un, KCNA, The, Lotte, Radio Free Asia, Ministry of Unification, Kuala Lumpur, Unification Ministry, -nam, Thae, country, two, WorkersṔarty of Korea, ministry, government, Rodong Sinmun, last, Jeong Joon, inter-, Malaysian, Warmbier, Ri, South Koreans, Han, Kaesong Industrial Complex, Republic of Korea, year

(4) **Sewol ho** Seoul, South Korea, Yonhap, said, The, South Koreans, percent, Gyeonggi Province, Sewol, Kim Da, Busan, South Korean, year, Incheon, government, people, million, Daegu, Bak Se, also, number, Gangwon Province, country, last, Seoul Metropolitan Government, Gwangju, years, @, heraldcorp.com, A, ministry, city, National Election Commission, South Jeolla Province, police, Pohang, Gwanghwamun Square, ś, one, Mokpo, public, Jeju, North Chungcheong Province, English, OECD, Jindo

(5) **Donald Trump** US, North Korea, North, South Korea, Washington, Seoul, North Korean, Trump, Pyongyang, United States, Yonhap, Donald Trump, Korean, “, said, ś, ´´, China, missile, American, nuclear, South Korean, UN, Korea, Russia, Kim Jong, UN Security Council, ICBM, White House, The,

DPRK, Rex Tillerson, U.S., military, Japan, South, WASHINGTON, UNSC, Tillerson, ballistic, also, Pacific, Guam, sanctions, would, State Department, test

(6) **The next presidential candidate** Liberty Korea Party, Hwang, Democratic Party, Bareun Party, Ahn, Hwang Kyo, Hong, Peopleś Party, Park Geun, Democratic Party of Korea, Yoo, Saenuri Party, Ahn Cheol, Yonhap, People Party, party, National Assembly, Hong Joon, Saenuri, South Korea, Justice Party, presidential, Yoo Seong, Rep, election, Jo He, percent, Moon Jae, South Chungcheong, Realmeter, conservative, Constitution, Constitutional Court, said, opposition, Sim Sang, Ko, The, Macau, former, ruling, candidate, Lee Jae, Rep. Yoo Seong, Gallup Korea, political

(7) **Korea-China Relationship and DAPA** China, South Korea, Chinese, THAAD, South Korean, Beijing, Seoul, Yonhap, Terminal High Altitude Area Defense, Xi Jinping, Army, Navy, US, Air Force, Seongju, ś, system, Danish, Xi, said, deployment, defense, Defense Ministry, Asia, Ministry of National Defense, Vietnam, Taiwan, Lim, ´´, Wang Yi, North Gyeongsang Province, The, Denmark, “, Coast Guard, military, DAPA, two, Defense Acquisition Program Administration, ministry, countries, Lim Sung, Incheon International Airport

(8) **Elected Candidate Moon Jae-in** Moon, Moon Jae, Cheong Wa Dae, Seoul, National Assembly, Yonhap, Democratic Party, -in, ś, President, “, Roh Moo, Moon, Cabinet, ´´, Park Geun, Roh, Kim, said, presidential, Lee Nak, government, president, Bae Hyun, Wa Dae, Choi He, also, new, office, The, Chung, LKP, Lee Myung, Park Soo, meeting, I, Chung Sye, South Korean, Liberty Korea Party, Yoon, opposition, Jun, chief, former, Kim Dong, Supreme Court, Republic of Korea, South Korea, public

(9) **European Union** Korea, Korean, Seoul, Joel Lee, The Korea Herald, Koreans, France, Europe, French, British, European, Canada, I, Canadian, Germany, Britain, The, Kazakhstan, EU, Paris, European Union, Poland, said, –, German, Iran, India, Ock Hyun, By, London, Park Hyun, Pakistan, country, Asia, Embassy, Russia, UK, countries, English, FTA, Africa, Morocco, African, Norway, Italian, Afghanistan

(10) **ASEAN and USFK** Malaysia, UN, Ban, Malaysian, ASEAN, Indonesia, Ban Ki, Philippines, Indonesian, Australia, Korea, Singapore, Kuala Lumpur, Vietnam, Southeast Asian, Korean, Asian, Thailand, United Nations, USFK, Vietnamese, NATO, Manila, Ukraine, Association of Southeast Asian Nations, Georgia, Park Young, Asia, Iraq, Seoul, Kuala Lumpur International Airport, Jakarta, said, Australian, Southeast Asia, Philippine, Cambodia, Kang Chol, Myanmar, Lunar New Year, The, Pacific, Ri Jong, New York, Kim Young, National Pension Service, Lao

A.4 On-issue yearly result: MERS

(1) **2015.01.** Chaebol scions promoted to executives at young age
(2) **2015.02.** Umbrella union set to launch general strike in April
(3) **2015.03.** Park calls for compromise on labor, pension reforms
(4) **2015.04.** Rift prevents closure on ferry disaster
(5) **2015.05.** Presidential office blames parties for failed pension bill
(6) **2015.06.** S. Korea confirms 3 more MERS cases, total rises to 18
(7) **2015.07.** S. Korea reports no new MERS cases for 17th day
(8) **2015.08.** Park gives appointment letter to new health minister
(9) **2015.09.** Moon stakes leadership on party reform
(10) **2015.10.** 61 isolated after last MERS patient rediagnosed
(11) **2015.11.** Rival parties split over violence at protest rally
(12) **2015.12.** Police to ban another massive rally

A.5 On-issue yearly result: President Park's Scandal

(1) **2016.01.** Lawmaker gets 16-month jail term for receiving illegal political
(2) **2016.02.** Park fills spy agency's key posts with North Korea experts
(3) **2016.03.** Executive of national swimming body arrested over alleged embezzlement
(4) **2016.04.** Ex-chief of umbrella labor union indicted over alleged illegal rallies
(5) **2016.05.** Former senior prosecutor summoned over lobbying scandal
(6) **2016.06.** Former senior prosecutor arrested for lobbying scandal
(7) **2016.07.** CJ Group chief undergoes surgery to remove lung tumor
(8) **2016.08.** Police clear Maestro Chung of embezzlement allegations
(9) **2016.09.** Cheong Wa Dae denies claims of top aide's illicit fundraising
(10) **2016.10.** Choi faces probe over influence-peddling scandal
(11) **2016.11.** Presidential office says Park will follow whatever decision parliament makes on her fate
(12) **2016.12.** Park Geun-hye impeachment explained

A.6 Off-issue result: MERS

(1) **event 0**
who Korean Air
what heiress gets 1 year
when Thursday
where Seoul
why Korean Air
how , former vice president of Korean Air , to one

(2) **event 1**
who President Park
what welcomes labor reform deal
when Tuesday
where unknown
why President Park Geun-hye on Tuesday
how “ tough ” decision to compromise on reform measures that

(3) **event 2**
who S. Korea
what reports no new MERS cases

when the day before
where S. Korea
why S. Korea
how no new MERS cases for 14th day .

(4) **event 3**
who Minimum wage
what declared despite resistance
when Wednesday
where South Korea
why Minimum wage
how The South Korean government Wednesday announced next year 's minimum

(5) **event 4**
who the chief of the Korea Confederation of Trade Unions
what walked out of the temple
when 11:20 a.m.
where Seoul
why police
how Unions voluntarily walked out of the temple in central Seoul

A.7 Off-issue result: President Park's Scandal

(1) **event 0**
who she
what repeatedly rejected to appear at a parliamentary hearing.Members
when Monday
where Seoul
why Lawmakers
how she repeatedly rejected to appear at a parliamentary hearing.Members of

(2) **event 1**
who Independent counsel Park Young-soo
what faces the daunting task
when Dec. 11
where Mir
why They
how Can independent counsel untangle Choi scandal ?

(3) **event 2**
who President Park Geun-hye
what is impeached
when Saturday
where Gwanghwamun Square
why because of the people around me . ”
how only halfway through ‘ .

(4) **event 3**
who A formal arrest warrant
what has been issued Thursday
when Thursday
where Grand Korea
why Choi
how 's longtime confidante , accused of collaborating with a presidential

(5) **event 4**
who Office
what give the person
when 7:30 a.m.
where Seoul
why Choi Soon-sil , the mysterious woman accused of interfering in state affairs using her decades-long relationship with President Park Geun-hye
how Choi Soon-sil returns ; Blue House ‘ raid ’ by

(6) **event 5**
who a special inspector
what can face a jail term
when Monday
where Seoul
why a special inspector
how Special inspector Lee Seok-su , who had been tasked with