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
lovelife@kaist.ac.kr

ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Jihee Park, Junseop Ji, and Soyoung Yoon. 2019. CS474 term project paper. In *CS474 Term project report* . KAIST, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

2 OVERVIEW

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3 TREND ANALYSIS

0. Data Preprocessing 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. 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. 3. Libraries Used For this project, we used pandas and gensim python libraries.

1. Experiments 0. Idea 0-0. Main idea and 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 this project. A recent study(2018) on Topic Modeling shows that Topic Quality improves when Named Entities are promoted.[ref: https://www.aclweb.org/anthology/P18-2040.pdf] 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 flexibily, but results conducted by this paper shows that setting alpha as 10 showed the best results. We take advantage of this paper and implement Named Entity Promoted Topic Modeling done by LDA.

1. Data Tokenization 1-1. Lemmatization is 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 function was taken from gensim.utils.lemmatize. “‘res.append(lemmatize(raw$_text,$ $stopwords))$“‘$Butafterwedolemmatization, removestopwords, andtokenizethedat$

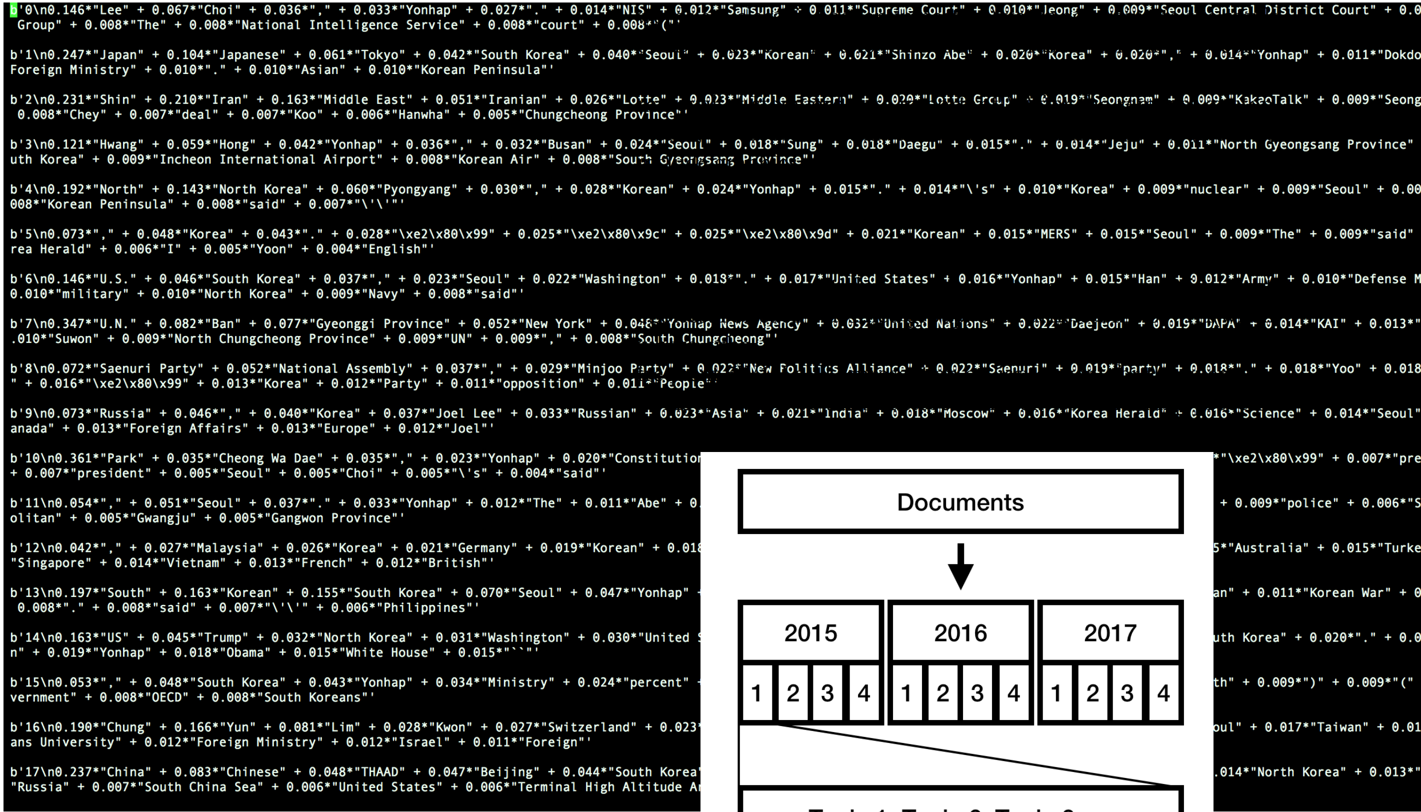
For these 3 reasons, we decided to NOT apply lemmatization for tokenization, because lemmatization lose so much information about the original text and disrupts the NER system’s ability to detect Named Entities properly. We decided to just do POS tagging and then do NER. We just used word$_tokenizefromnltk.tokenize.$

2. Extract NER By using ne$_chunkfromnltkandpos_tagfromnltk.tag, weextracte$ $wordinformationofNamedEntitiesotherthanjustclassifyingwhetherawordisand$ $wordNamedEntitiesandmulti-wordnamedentitiesseparately.Asaresult, NERand$ $wordextractionofNERarebothprocessed.$

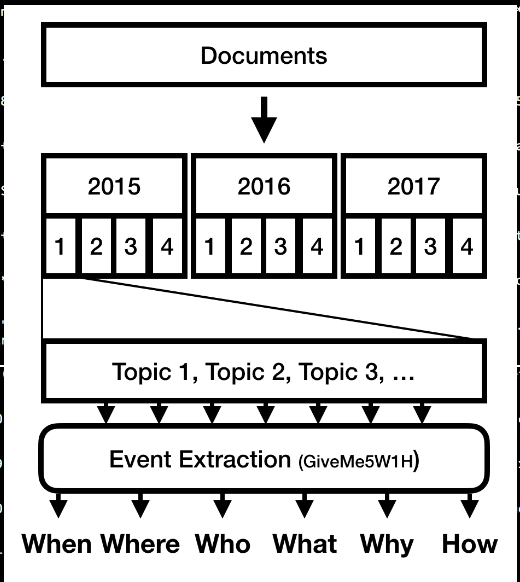
below figure is the topic modeling result(of all time lengths from 2015 to 2017) WITH NER Promoting and WITHOUT NER Promoting. We can see the difference between those two results, and we can

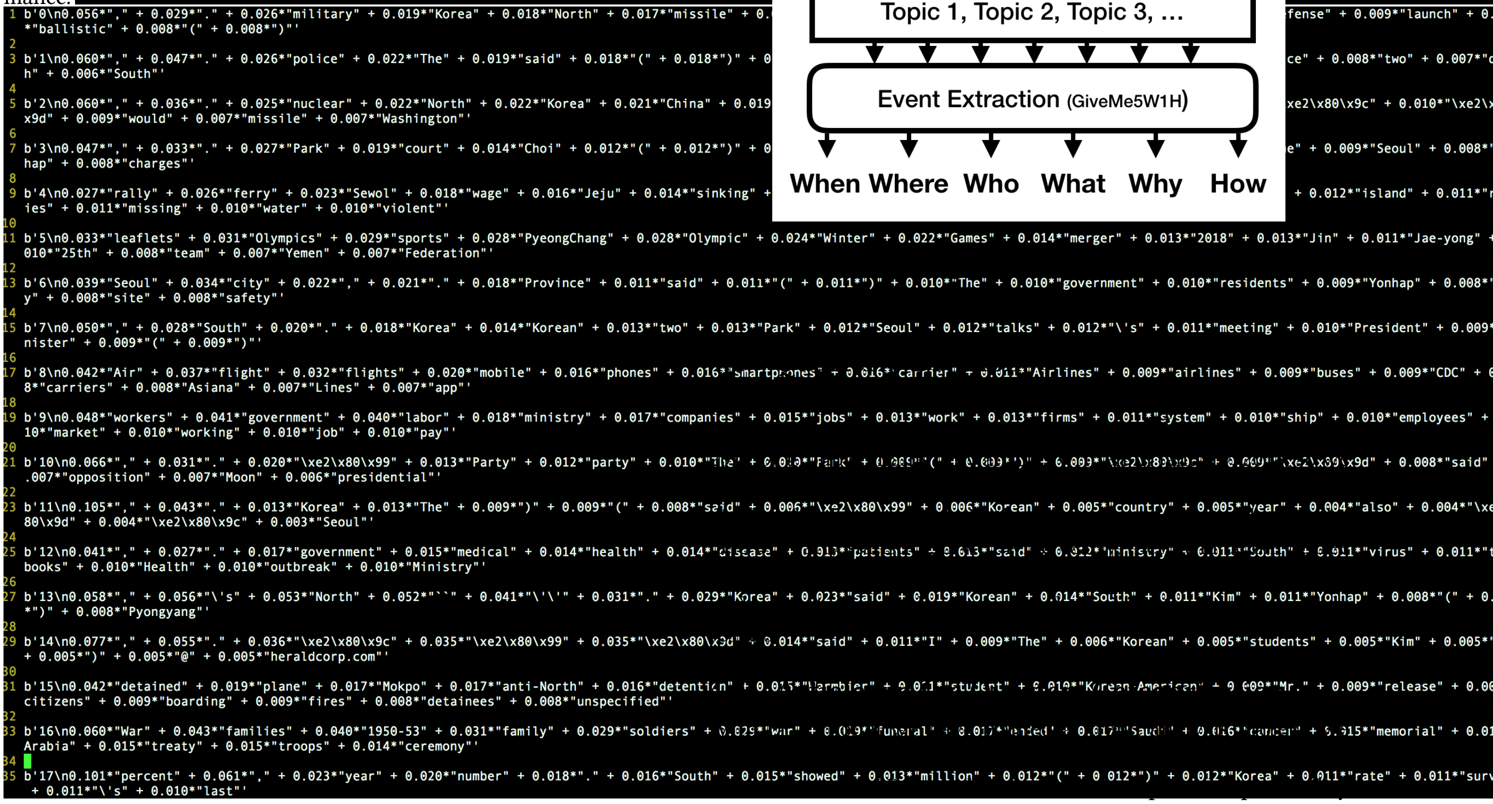
CS474 Term Project, 2019 Fall, KAIST
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

conclude topic modeling with NER promoting shows better performance.



```
b'0\n0.146*"Lee" + 0.067*"Choi" + 0.036*"," + 0.033*"Yonhap" + 0.027*"." + 0.014*"NIS" + 0.012*"Samsung" + 0.011*"Supreme Court" + 0.010*"Jeong" + 0.009*"Seoul Central District Court" + 0.0
 Group" + 0.008*"The" + 0.008*"National Intelligence Service" + 0.008*"court" + 0.008*"("'

b'1\n0.247*"Japan" + 0.104*"Japanese" + 0.061*"Tokyo" + 0.042*"South Korea" + 0.040*"Seoul" + 0.023*"Korean" + 0.021*"Shinzo Abe" + 0.020*"Korea" + 0.020*"," + 0.014*"Yonhap" + 0.011*"Dokdo
Foreign Ministry" + 0.010*"." + 0.010*"Asian" + 0.010*"Korean Peninsula"'

b'2\n0.231*"Shin" + 0.210*"Iran" + 0.163*"Middle East" + 0.051*"Iranian" + 0.026*"Lotte" + 0.023*"Middle Eastern" + 0.020*"Lotte Group" + 0.019*"Seongnam" + 0.009*"KakaoTalk" + 0.009*"Seong
 0.008*"Chey" + 0.007*"deal" + 0.007*"Koo" + 0.006*"Hanwha" + 0.005*"Chungcheong Province"'

b'3\n0.121*"Hwang" + 0.059*"Hong" + 0.042*"Yonhap" + 0.036*"," + 0.032*"Busan" + 0.024*"Seoul" + 0.018*"Sung" + 0.018*"Daegu" + 0.015*"." + 0.014*"Jeju" + 0.011*"North Gyeongsang Province"
uth Korea" + 0.009*"Incheon International Airport" + 0.008*"Korean Air" + 0.008*"South Gyeongsang Province"'

b'4\n0.192*"North" + 0.143*"North Korea" + 0.060*"Pyongyang" + 0.030*"," + 0.028*"Korean" + 0.024*"Yonhap" + 0.015*"." + 0.014*"\'s" + 0.010*"Korea" + 0.009*"nuclear" + 0.009*"Seoul" + 0.00
008*"Korean Peninsula" + 0.008*"said" + 0.007*"\'\'"'

b'5\n0.073*"," + 0.048*"Korea" + 0.043*"." + 0.028*"\xe2\x80\x99" + 0.025*"\xe2\x80\x9c" + 0.025*"\xe2\x80\x9d" + 0.021*"Korean" + 0.015*"MERS" + 0.015*"Seoul" + 0.009*"The" + 0.009*"said"
rea Herald" + 0.006*"I" + 0.005*"Yoon" + 0.004*"English"'

b'6\n0.146*"U.S." + 0.046*"South Korea" + 0.037*"," + 0.023*"Seoul" + 0.022*"Washington" + 0.018*"." + 0.017*"United States" + 0.016*"Yonhap" + 0.015*"Han" + 0.012*"Army" + 0.010*"Defense M
0.010*"military" + 0.010*"North Korea" + 0.009*"Navy" + 0.008*"said"'

b'7\n0.347*"U.N." + 0.082*"Ban" + 0.077*"Gyeonggi Province" + 0.052*"New York" + 0.048*"Yonhap News Agency" + 0.032*"United Nations" + 0.022*"Daejeon" + 0.019*"DAPA" + 0.014*"KAI" + 0.013*"
.010*"Suwon" + 0.009*"North Chungcheong Province" + 0.009*"UN" + 0.009*"," + 0.008*"South Chungcheong"'

b'8\n0.072*"Saenuri Party" + 0.052*"National Assembly" + 0.037*"," + 0.029*"Minjoo Party" + 0.022*"New Politics Alliance" + 0.022*"Saenuri" + 0.019*"party" + 0.018*"." + 0.018*"Yoo" + 0.018
" + 0.016*"\xe2\x80\x99" + 0.013*"Korea" + 0.012*"Party" + 0.011*"opposition" + 0.011*"People"'

b'9\n0.073*"Russia" + 0.046*"," + 0.040*"Korea" + 0.037*"Joel Lee" + 0.033*"Russian" + 0.023*"Asia" + 0.021*"India" + 0.018*"Moscow" + 0.016*"Korea Herald" + 0.016*"Science" + 0.014*"Seoul"
anada" + 0.013*"Foreign Affairs" + 0.013*"Europe" + 0.012*"Joel"'

b'10\n0.361*"Park" + 0.035*"Cheong Wa Dae" + 0.035*"," + 0.023*"Yonhap" + 0.020*"Constitution                  *"\xe2\x80\x99" + 0.007*"pre
+ 0.007*"president" + 0.005*"Seoul" + 0.005*"Choi" + 0.005*"\'s" + 0.004*"said"'

b'11\n0.054*"," + 0.051*"Seoul" + 0.037*"." + 0.033*"Yonhap" + 0.012*"The" + 0.011*"Abe" + 0.                  + 0.009*"police" + 0.006*"S
olitan" + 0.005*"Gwangju" + 0.005*"Gangwon Province"'

b'12\n0.042*"," + 0.027*"Malaysia" + 0.026*"Korea" + 0.021*"Germany" + 0.019*"Korean" + 0.018                  5*"Australia" + 0.015*"Turke
"Singapore" + 0.014*"Vietnam" + 0.013*"French" + 0.012*"British"'

b'13\n0.197*"South" + 0.163*"Korean" + 0.155*"South Korea" + 0.070*"Seoul" + 0.047*"Yonhap" +                  an" + 0.011*"Korean War" + 0
 0.008*"." + 0.008*"said" + 0.007*"\'\'" + 0.006*"Philippines"'

b'14\n0.163*"US" + 0.045*"Trump" + 0.032*"North Korea" + 0.031*"Washington" + 0.030*"United S                  uth Korea" + 0.020*"." + 0.0
n" + 0.019*"Yonhap" + 0.018*"Obama" + 0.015*"White House" + 0.015*"``"'

b'15\n0.053*"," + 0.048*"South Korea" + 0.043*"Yonhap" + 0.034*"Ministry" + 0.024*"percent" +                  th" + 0.009*")" + 0.009*"("
vernment" + 0.008*"OECD" + 0.008*"South Koreans"'

b'16\n0.190*"Chung" + 0.166*"Yun" + 0.081*"Lim" + 0.028*"Kwon" + 0.027*"Switzerland" + 0.023*                  oul" + 0.017*"Taiwan" + 0.01
ans University" + 0.012*"Foreign Ministry" + 0.012*"Israel" + 0.011*"Foreign"'

b'17\n0.237*"China" + 0.083*"Chinese" + 0.048*"THAAD" + 0.047*"Beijing" + 0.044*"South Korea"                  .014*"North Korea" + 0.013*"
"Russia" + 0.007*"South China Sea" + 0.006*"United States" + 0.006*"Terminal High Altitude Ar
```





```
b'0\n0.056*"," + 0.029*"." + 0.026*"military" + 0.019*"Korea" + 0.018*"North" + 0.017*"missile" + 0.                  fense" + 0.009*"launch" + 0.
*"ballistic" + 0.008*"(" + 0.008*")"'

b'1\n0.060*"," + 0.047*"." + 0.026*"police" + 0.022*"The" + 0.019*"said" + 0.018*"(" + 0.018*")" + 0                  ce" + 0.008*"two" + 0.007*"c
h" + 0.006*"South"'

b'2\n0.060*"," + 0.036*"." + 0.025*"nuclear" + 0.022*"North" + 0.022*"Korea" + 0.021*"China" + 0.019                  xe2\x80\x9c" + 0.010*"\xe2\x
x9d" + 0.009*"would" + 0.007*"missile" + 0.007*"Washington"'

b'3\n0.047*"," + 0.033*"." + 0.027*"Park" + 0.019*"court" + 0.014*"Choi" + 0.012*"(" + 0.012*")" + 0                  e" + 0.009*"Seoul" + 0.008*"
hap" + 0.008*"charges"'

b'4\n0.027*"rally" + 0.026*"ferry" + 0.023*"Sewol" + 0.018*"wage" + 0.016*"Jeju" + 0.014*"sinking" +                  + 0.012*"island" + 0.011*"r
ies" + 0.011*"missing" + 0.010*"water" + 0.010*"violent"'

b'5\n0.033*"leaflets" + 0.031*"Olympics" + 0.029*"sports" + 0.028*"PyeongChang" + 0.028*"Olympic" + 0.024*"Winter" + 0.022*"Games" + 0.014*"merger" + 0.013*"2018" + 0.013*"Jin" + 0.011*"Jae-yong" +
010*"25th" + 0.008*"team" + 0.007*"Yemen" + 0.007*"Federation"'

b'6\n0.039*"Seoul" + 0.034*"city" + 0.022*"," + 0.021*"." + 0.018*"Province" + 0.011*"said" + 0.011*"(" + 0.011*")" + 0.010*"The" + 0.010*"government" + 0.010*"residents" + 0.009*"Yonhap" + 0.008*"
y" + 0.008*"site" + 0.008*"safety"'

b'7\n0.050*"," + 0.028*"South" + 0.020*"." + 0.018*"Korea" + 0.014*"Korean" + 0.013*"two" + 0.013*"Park" + 0.012*"Seoul" + 0.012*"talks" + 0.012*"\'s" + 0.011*"meeting" + 0.010*"President" + 0.009*
nister" + 0.009*"(" + 0.009*")"'

b'8\n0.042*"Air" + 0.037*"flight" + 0.032*"flights" + 0.020*"mobile" + 0.016*"phones" + 0.016*"smartphones" + 0.016*"carrier" + 0.011*"Airlines" + 0.009*"airlines" + 0.009*"buses" + 0.009*"CDC" + 0
8*"carriers" + 0.008*"Asiana" + 0.007*"Lines" + 0.007*"app"'

b'9\n0.048*"workers" + 0.041*"government" + 0.040*"labor" + 0.018*"ministry" + 0.017*"companies" + 0.015*"jobs" + 0.013*"work" + 0.013*"firms" + 0.011*"system" + 0.010*"ship" + 0.010*"employees" +
10*"market" + 0.010*"working" + 0.010*"job" + 0.010*"pay"'

b'10\n0.066*"," + 0.031*"." + 0.020*"\xe2\x80\x99" + 0.013*"Party" + 0.012*"party" + 0.010*"The" + 0.010*"Park" + 0.009*"(" + 0.009*")" + 0.009*"\xe2\x80\x9c" + 0.009*"\xe2\x80\x9d" + 0.008*"said"
.007*"opposition" + 0.007*"Moon" + 0.006*"presidential"'

b'11\n0.105*"," + 0.043*"." + 0.013*"Korea" + 0.013*"The" + 0.009*")" + 0.009*"(" + 0.008*"said" + 0.006*"\xe2\x80\x99" + 0 006*"Korean" + 0.005*"country" + 0.005*"year" + 0.004*"also" + 0.004*"\xe
80\x9d" + 0.004*"\xe2\x80\x9c" + 0.003*"Seoul"'

b'12\n0.041*"," + 0.027*"." + 0.017*"government" + 0.015*"medical" + 0.014*"health" + 0.014*"disease" + 0.013*"patients" + 0.013*"said" + 0.012*"ministry" + 0.011*"South" + 0.011*"virus" + 0.011*"t
books" + 0.010*"Health" + 0.010*"outbreak" + 0.010*"Ministry"'

b'13\n0.058*"," + 0.056*"\'s" + 0.053*"North" + 0.052*"``" + 0.041*"\'\'" + 0.031*"." + 0.029*"Korea" + 0.023*"said" + 0.019*"Korean" + 0.014*"South" + 0.011*"Kim" + 0.011*"Yonhap" + 0.008*"(" + 0.
*")" + 0.008*"Pyongyang"'

b'14\n0.077*"," + 0.055*"." + 0.036*"\xe2\x80\x9c" + 0.035*"\xe2\x80\x99" + 0.035*"\xe2\x80\x9d" + 0.014*"said" + 0.011*"I" + 0.009*"The" + 0.006*"Korean" + 0.005*"students" + 0.005*"Kim" + 0.005*"
+ 0.005*")" + 0.005*"@" + 0.005*"heraldcorp.com"'

b'15\n0.042*"detained" + 0.019*"plane" + 0.017*"Mokpo" + 0.017*"anti-North" + 0.016*"detention" + 0.015*"Warmbier" + 0.011*"student" + 0.010*"Korean-American" + 0 009*"Mr." + 0.009*"release" + 0.00
citizens" + 0.009*"boarding" + 0.009*"fires" + 0.008*"detainees" + 0.008*"unspecified"'

b'16\n0.060*"War" + 0.043*"families" + 0.040*"1950-53" + 0.031*"family" + 0.029*"soldiers" + 0.029*"war" + 0.019*"funeral" + 0.017*"ended" + 0.017*"Saudi" + 0.016*"cancer" + 0.015*"memorial" + 0.01
Arabia" + 0.015*"treaty" + 0.015*"troops" + 0.014*"ceremony"'

b'17\n0.101*"percent" + 0.061*"," + 0.023*"year" + 0.020*"number" + 0.018*"." + 0.016*"South" + 0.015*"showed" + 0.013*"million" + 0.012*"(" + 0 012*")" + 0.012*"Korea" + 0.011*"rate" + 0.011*"surv
 + 0.011*"\'s" + 0.010*"last"'
```

3. Do LDA At first try, we ran LDA . But we found out that stop-

for each group. We choose the group with maximum value, and assign the document to the group. So, for each quarter, there are 20 classified groups of news articles.

4.3 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". 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 an event.

5 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 $\mathbb{T}_a \in \mathcal{P}(\mathbb{D})$.

5.1 BoW extractuion

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 $\phi(\mathbb{T}_a) = \{b(d) | d \in \mathbb{T}_a\}$

5.2 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 [?].

Our desired result is that b with euclidean distance successfuly 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 5.1. b has approximation of ϕ with bound K, ϵ iff there exists an Lipshitz continuous π with K that $d^\#(\pi(b(d)), \phi(d)) \le \epsilon$.

5.3 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 5.2. e is Lipschitz continuous with K if and only if $e^\#(e(d_1), e(d_2)) \le K d^\#(d_1, d_2)$.

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 approximation of semantics with bounded error, then there also exists approximation of event with bounded error.

Theorem 5.3. *b has approximation of ϕ with bound K, ϵ, then there exists $\pi_e : \mathbb{R}^{n+1} \to E^\#$ s.t. $e^\#(\pi_e(b(d)), e(\phi(d))) \le K_e \cdot \epsilon$. (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.

5.4 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 5.4. *if b has approximation of $e \cdot \phi$ with bound K, ϵ, then $e^\#(e \cdot \phi(d_1), e \cdot \phi(d_2)) \le 2 \cdot \epsilon + K||b(d_1) - b(d_2)||$.*

Proof.

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

□

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

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

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 $\approx_{E^{\#},2\cdot\epsilon+K\cdot\delta}\cdot$

Definition 5.6 (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.

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 5.7 (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 relevent description of event is extracted from each partiton.

5.5 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 docuemnts, 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 bewtween description of similar/same meaning.

For example, if one document describe the event happens "one day after of 12/7", and there are another document describe the event was happend "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 5.8 (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[?] to extract description of event.

5.6 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.

6 EVALUATION

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7 CONCLUSION

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.