

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, 2 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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>

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 flexibly, 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))” “But after we do lemmatization, remove stopwords, and tokenize the data

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_tokenize` from `nltk.tokenize`.

2. Extract NER By using `ner_chunk` from `nltk` and `pos_tag` from `nltk.tag`, we extract word information of Named Entities other than just classifying whether a word is a Named Entity and multi-word named entities separately. As a result, NER and 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 perfor-

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

175

0.008**"The" + 0.008**"National Intelligence Service" + 0.008**"court" + 0.008**(""

b'1\0.247**"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\0.231**"Shin" + 0.210**"Iran" + 0.163**"Middle East" + 0.051**"Iranian" + 0.026**"Lotte" + 0.013**"Middle Eastern" + 0.020**"Lotte Group" + 0.019**"Seongnam" + 0.009**"Kakaotalk" + 0.009**"Seoul 0.008**"Chey" + 0.007**"deal" + 0.007**"Koo" + 0.006**"Hanwha" + 0.005**"Chungcheong Province"

b'3\0.121**"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\0.192**"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\0.073**", " + 0.048**"Korea" + 0.043**", " + 0.028**"xe2\x80\x99" + 0.025**"xe2\x80\x99" + 0.025**"xe2\x80\x99" + 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\0.146**"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\0.347**"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**"DAFA" + 0.014**"KAI" + 0.013** ,010**"Suwon" + 0.009**"North Chungcheong Province" + 0.009**"UN" + 0.009**", " + 0.008**"South Chungcheong"

b'8\0.072**"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\0.073**"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\0.361**"Park" + 0.035**"Cheong Wa Dae" + 0.035**", " + 0.023**"Yonhap" + 0.020**"Constitutional Court" + 0.015**", " + 0.012**"President" + 0.011**"Geun-hye" + 0.008**"xe2\x80\x99" + 0.007**"pre + 0.007**"president" + 0.005**"Seoul" + 0.005**"Choi" + 0.005**"\'s" + 0.004**"said"

b'11\0.054**", " + 0.051**"Seoul" + 0.037**", " + 0.033**"Yonhap" + 0.012**"The" + 0.011**"Abe" + 0.011**"said" + 0.011**"Sewol" + 0.010**(" + 0.010**") + 0.009**"Incheon" + 0.009**"police" + 0.006**"S olitan" + 0.005**"Gwangju" + 0.005**"Gangwon Province"

b'12\0.042**", " + 0.027**"Malaysia" + 0.026**"Korea" + 0.021**"Germany" + 0.019**"Korean" + 0.018**"Paris" + 0.017**"France" + 0.016**"Indonesia" + 0.016**"ASEAN" + 0.015**"Australia" + 0.015**"Turke "Singapore" + 0.014**"Vietnam" + 0.013**"French" + 0.012**"British"

b'13\0.197**"South" + 0.163**"Korean" + 0.155**"South Korea" + 0.070**"Seoul" + 0.047**"Yonhap" + 0.019**", " + 0.017**"South Koreans" + 0.011**"\'s" + 0.011**"South Korean" + 0.011**"Korean War" + 0 0.008**", " + 0.008**"said" + 0.007**"\'\'\' + 0.006**"Philippines"

b'14\0.163**"US" + 0.045**"Trump" + 0.032**"North Korea" + 0.031**"Washington" + 0.030**"United States" + 0.020**", " + 0.024**"Donald Trump" + 0.023**"Kang" + 0.020**"South Korea" + 0.020**", " + 0.0 n" + 0.019**"Yonhap" + 0.018**"Obama" + 0.015**"White House" + 0.015**""

b'15\0.053**", " + 0.048**"South Korea" + 0.043**"Yonhap" + 0.034**"Ministry" + 0.024**"percent" + 0.023**", " + 0.011**"said" + 0.010**"The" + 0.009**"year" + 0.009**"Health" + 0.009**") + 0.009**(" vernment" + 0.008**"OECD" + 0.008**"South Koreans"

b'16\0.190**"Chung" + 0.166**"Yun" + 0.081**"Lim" + 0.028**"Kwon" + 0.027**"Switzerland" + 0.023**"Cuba" + 0.022**"Laos" + 0.019**"Gangnam" + 0.019**"Norway" + 0.018**"Seoul" + 0.017**"Taiwan" + 0.01 ans University" + 0.012**"Foreign Ministry" + 0.012**"Israel" + 0.011**"Foreign"

b'17\0.237**"China" + 0.083**"Chinese" + 0.048**"THAAD" + 0.047**"Beijing" + 0.044**"South Korea" + 0.031**"U.S." + 0.031**"Seoul" + 0.022**", " + 0.015**"Washington" + 0.014**"North Korea" + 0.013** "Russia" + 0.007**"South China Sea" + 0.006**"United States" + 0.006**"Terminal High Altitude Area"

mance,

1 b'0\0.056**", " + 0.029**", " + 0.026**"military" + 0.019**"Korea" + 0.018**"North" + 0.017**"missile" + 0.017**"South" + 0.012**"said" + 0.012**"The" + 0.010**"Korean" + 0.009**"defense" + 0.009**"launch" + 0. 2 **"ballistic" + 0.008**(" + 0.008**")'

3 h" + 0.006**"South"

4

5 b'2\0.060**", " + 0.036**", " + 0.025**"nuclear" + 0.022**"North" + 0.022**"Korea" + 0.021**"China" + 0.019**"U.S." + 0.014**"xe2\x80\x99" + 0.011**"JS" + 0.011**"said" + 0.010**"xe2\x80\x99" + 0.010**"xe2\x80\x99" + 0.009**"would" + 0.007**"missile" + 0.007**"Washington"

6

7 b'3\0.047**", " + 0.033**", " + 0.027**"Park" + 0.019**"court" + 0.014**"Choi" + 0.012**(" + 0.012**") + 0.011**"Court" + 0.011**"former" + 0.010**"Lee" + 0.010**"\'s" + 0.009**"The" + 0.009**"Seoul" + 0.008** hap" + 0.008**"charges"

8

9 b'4\0.027**"rally" + 0.026**"ferry" + 0.023**"Sewol" + 0.018**"wage" + 0.016**"Jeju" + 0.014**"sinking" + 0.014**"protest" + 0.013**"police" + 0.012**"people" + 0.012**"disaster" + 0.012**"island" + 0.011**" ies" + 0.011**"missing" + 0.010**"water" + 0.010**"violent"

10

11 b'5\0.033**"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" + 0 010**"25th" + 0.008**"team" + 0.007**"Yemen" + 0.007**"Federation"

12

13 b'6\0.039**"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"

14

15 b'7\0.050**", " + 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**")'

16

17 b'8\0.042**"Air" + 0.037**"flight" + 0.032**"flights" + 0.020**"mobile" + 0.016**"phones" + 0.016**"smartphones" + 0.016**"carrier" + 0.011**"Air-lines" + 0.009**"air-lines" + 0.009**"buses" + 0.009**"CDC" + 0 8**"carriers" + 0.008**"Asiana" + 0.007**"Lines" + 0.007**"app"

18

19 b'9\0.048**"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"

20

21 b'10\0.066**", " + 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\x99" + 0.009**"xe2\x80\x99" + 0.008**"said" ,007**"opposition" + 0.007**"Moon" + 0.006**"presidential"

22

23 b'11\0.105**", " + 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\x99" + 0.004**"xe2\x80\x99" + 0.003**"Seoul"

24

25 b'12\0.041**", " + 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**" books" + 0.010**"Health" + 0.010**"outbreak" + 0.010**"Ministry"

26

27 b'13\0.058**", " + 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"

28

29 b'14\0.077**", " + 0.055**", " + 0.036**"xe2\x80\x99" + 0.035**"xe2\x80\x99" + 0.035**"xe2\x80\x99" + 0.014**"said" + 0.011**"I" + 0.009**"The" + 0.006**"Korean" + 0.005**"students" + 0.005**"Kim" + 0.005**" + 0.005**") + 0.005**"heraldcorp.com"

30

31 b'15\0.042**"detained" + 0.019**"plane" + 0.017**"Mokpo" + 0.017**"anti-North" + 0.016**"detention" + 0.015**"Warmbler" + 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"

32

33 b'16\0.060**"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"

34

35 b'17\0.101**"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"

231

232