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$_{t}ext,$ $stopwords))$“‘$Butafterwedolemmatization, removestopwords, andtokenizethedat$

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$_{t}okenizefromnltk.tokenize.$

2. Extract NER By using ne$_{c}hunkfromnltkandpos_{t}agfromnltk.tag, weextracte$ $wordinformationofNamedEntitiesotherthanjustclassifyingwhetherawordisanc$ $wordNamedEntitiesandmulti-wordnamedentitiesseparately.Asaresult, NERanc$ $wordextractionofNERarebothprocessed.$

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

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

can conclude topic modeling with NER promoting shows better performance. after_ner.png before_ner.png 3. Do LDA At first try, we ran LDA . But 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)were done. (The timing of removal of stopwords are important!) After the removal of stopwords, we could see that the result were much better. (Need to include graphics or charts)

4 ON-ISSUE TRACKING

4.1 Method

(1) **Classification** We perform classification based on the LDA model that had been trained before.
(2) **Event Extraction** Woah! Giveme5W1H! Very trash!

4.2 Result

5 OFF-ISSUE TRACKING

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.

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.