

english



จุฬาลงกรณ์มหาวิทยาลัย  
**CHULALONGKORN UNIVERSITY**

วิธีการตัดหน่วยใหม่โดยอิงการปรากฏร่วมเพื่อใช้ในแบบจำลองการแจกแจงหัวข้อด้วยการ  
แจกแจงดีริชเลแฝง



สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต

สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2564

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

COLLOCATION-BASED RETOKENIZATION METHODS FOR LATENT  
DIRICHLET ALLOCATION TOPIC MODELS

Miss Jin Cheevaprawatdomrong



An Independent Study Submitted in Partial Fulfillment of the Requirements  
for the Degree of Master of Arts Program in Linguistics

Department of Linguistics

Faculty of Arts

Chulalongkorn University

Academic Year 2021

Copyright of Chulalongkorn University

Independent Study Title COLLOCATION-BASED RETOKENIZATION METHODS FOR LATENT DIRICHLET ALLOCATION TOPIC MODELS

By Miss Jin Cheevaprawatdomrong

Field of Study Linguistics

Independent Study Advisor Assistant Professor Attapol Thamrongrattanarit, Ph.D.

---

Accepted by the Faculty of Arts, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

Dean of the Faculty of Arts

.....  
(Associate Professor Suradech Chotiudompant, Ph.D.)

INDEPENDENT STUDY COMMITTEE

..... Chairman

(Assistant Professor Theeraporn Ratitamkul, Ph.D.)

..... Independent Study Advisor

(Assistant Professor Attapol Thamrongrattanarit, Ph.D.)

..... Examiner

(Associate Professor Wirote Aroonmanakun, Ph.D.)

จินต์ ชิวประวัติดำรงค์: วิธีการตัดหน่วยใหม่โดยอิงการปรากฏร่วมเพื่อใช้ในแบบจำลองการแจกหัวข้อด้วยการแจกแจงดีริชเลแฝง. (COLLOCATION-BASED RETOKENIZATION METHODS FOR LATENT DIRICHLET ALLOCATION TOPIC MODELS) อ.ที่ปริกษาสารนิพนธ์หลัก : รศ. ดร. อรรถพล ชำรงรัตนฤทธิ์, 80 หน้า.

การจัดสรรดีริชเลแฝงสามารถค้นพบหัวข้อต่างๆที่แฝงอยู่ในเอกสารโดยใช้คำเป็นสิ่ง ที่ป้อนเข้า งานวิจัยที่ผ่านมาแสดงว่าการรวมคำเป็นคำปรากฏร่วมสามารถทำให้หัวข้อที่ได้มี ความเชื่อมโยงกันมากขึ้นในภาษาอังกฤษ แต่ยังคงมีคำถามว่าวิธีใดเป็นวิธีที่ดีที่สุดที่จะรวม คำเข้าด้วยกัน โดยเฉพาะอย่างยิ่งในภาษาที่ไม่มีสัญลักษณ์แบ่งคำที่ชัดเจนอย่างภาษาจีนและ ภาษาไทย ผู้ดำเนินงานวิจัยได้เปรียบเทียบวิธี การทดสอบไคสแควร์ สถิติทดสอบที และ ความถี่ และแสดงว่าการรวมคำที่ป้อนเข้าด้วยวิธีที่เหมาะสมจะสามารถทำให้ความเหมาะสม กับข้อมูลของแบบจำลอง (goodness of fit) และความเชื่อมโยงกันของหัวข้อของแบบจำลอง ดีขึ้น

จุฬาลงกรณ์มหาวิทยาลัย  
CHULALONGKORN UNIVERSITY

ภาควิชา	ภาษาศาสตร์	ลายมือชื่อนิสิต	.....
สาขาวิชา	ภาษาศาสตร์	ลายมือชื่อ อ.ที่ปริกษาหลัก	.....
ปีการศึกษา	2564		

## 6382008322: MAJOR LINGUISTICS

KEYWORDS: LATENT DIRICHLET ALLOCATION / COLLOCATION / TOKENIZATION

JIN CHEEVAPRAWATDOMRONG : COLLOCATION-BASED RETOKENIZATION METHODS FOR LATENT DIRICHLET ALLOCATION TOPIC MODELS. ADVISOR : ASST PROF ATTAPOL THAMRONGRATTANARIT, Ph.D., 80 pp.

Latent Dirichlet Allocation (LDA) discovers hidden themes in documents by using words as input. Past studies show that merging the words into collocation improves topic coherence in English. However, there are still questions about the best merging strategies, especially in the languages without clear word boundaries, such as Thai and Chinese. We compare chi-squared measure, *t*-statistics, and raw frequency strategies, and show that merging input tokens with appropriate strategies can improve the goodness of fit and topic coherence of the model.



จุฬาลงกรณ์มหาวิทยาลัย  
CHULALONGKORN UNIVERSITY

Department: Linguistics

Student's Signature .....

Field of Study: Linguistics

Advisor's Signature .....

Academic Year: 2021

## Acknowledgements

I would like to thank my advisor, Assistant Professor Attapol Thamrongrattanarit, Ph.D., for the chance to work with him on many interesting projects. Without such opportunities, I would not have started the research journey that I so much enjoy. I appreciate the time and effort he put into guiding me through the research and writing process. I am thankful to be his student, and I promise to pay it forward when I mentor my students if I could become a professor one day.

I would like to express my gratitude to all faculty members in the Linguistics Department for accepting me into this prestigious university. I gained much more understanding of linguistics while studying in their classes during these two years and became more confident in learning and conducting research further.

I would like to also thank my parents and my brother for their unconditional love and support, which help me to be tough to get through difficult times.

# CONTENTS

	<b>Page</b>
english	
<b>Abstract (Thai)</b> . . . . .	<b>iv</b>
<b>Abstract (English)</b> . . . . .	<b>v</b>
<b>Acknowledgements</b> . . . . .	<b>vi</b>
<b>Contents</b> . . . . .	<b>vii</b>
<b>List of Tables</b> . . . . .	<b>ix</b>
<b>List of Figures</b> . . . . .	<b>x</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
1.1 Latent Dirichlet Allocation . . . . .	1
1.2 Topics . . . . .	1
1.3 Collocations . . . . .	2
1.4 Tokenization . . . . .	2
1.5 Retokenization . . . . .	4
1.6 Objective and Hypothesis . . . . .	4
1.7 Contribution . . . . .	4
<b>2 Background and Literature Review</b> . . . . .	<b>6</b>
2.1 Latent Dirichlet Allocation . . . . .	6
2.2 Improving and Evaluating Topic Models . . . . .	10
2.3 Collocations . . . . .	12
2.4 Morphological Typology and Writing Systems . . . . .	21
<b>3 Our Proposed Method</b> . . . . .	<b>28</b>
3.1 Collocations as LDA Token . . . . .	28
3.2 Evaluation Metrics . . . . .	28



	viii Page
3.3 Experiments . . . . .	30
3.4 Results and Discussion . . . . .	33
<b>4 Conclusion . . . . .</b>	<b>41</b>
<b>References . . . . .</b>	<b>42</b>
<b>Appendix . . . . .</b>	<b>47</b>
<b>Appendix A Top Bigrams . . . . .</b>	<b>47</b>
<b>Appendix B Topic Keys . . . . .</b>	<b>55</b>



# LIST OF TABLES

Table	Page
english	
2.1 The nouns w occurring most often in the patterns “strong w” and “powerful w” from New York Times newswire (Manning and Schutze, 1999).	14
2.2 The number of occurrences of the word “some definition” and other words	18
3.1 The details of corpora we use in this study (Cheevaprawatdomrong et al., 2022) . . . . .	32
3.2 The percentage of overlapping merged tokens between two methods of retokenization computed on the retokenization training data. (Cheevaprawatdomrong et al., 2022) . . . . .	39
3.3 Normalized unigram log-likelihood per token (top) and Concatenation-based Embedding Silhouette (CBES) scores (bottom) for between the baseline and retokenization models: $\chi^2$ , texttitt, and raw frequency. (Cheevaprawatdomrong et al., 2022) . . . . .	40

# LIST OF FIGURES

Figure	Page
english	
1.1 Example of topics from restaurant reviews. . . . .	1
2.1 The intuition of the LDA model. (Blei, 2012) . . . . .	7
2.2 Example of topics from New York Times corpus. . . . .	7
2.3 The figure shows the relationship of random variables (Blei, 2012) . . . .	8
2.4 Histogram of the position of strong relative to three words from the New York Times newswire (Manning and Schutze, 1999) . . . . .	16
2.5 Chinese is an example of an analytic language . . . . .	22
2.6 Chinese is an example of logogram . . . . .	24
2.7 Japanese is an example of syllabary . . . . .	24
2.8 Hebrew is an example of abjad . . . . .	25
2.9 Thai is an example of abugida . . . . .	26
2.10 Korean is an example of featural system . . . . .	26
3.1 The topic on the right is a better topic . . . . .	29
3.2 The top 20 collocations from each strategy. (Cheevaprawatdomrong et al., 2022) . . . . .	37
3.3 PTLI improvement vs. merged percentage. (Cheevaprawatdomrong et al., 2022) . . . . .	38
3.4 CBES improvement vs. merged percentage. (Cheevaprawatdomrong et al., 2022) . . . . .	39
A.1 The top 50 collocations from English Wikipedia. (Cheevaprawatdom- rong et al., 2022) . . . . .	47
A.2 The top 50 collocations from German Wikipedia. (Cheevaprawatdom- rong et al., 2022) . . . . .	48
A.3 The top 50 collocations from Chinese Wikipedia. (Cheevaprawatdom- rong et al., 2022) . . . . .	49
A.4 The top 50 collocations from Japanese Wikipedia. (Cheevaprawat- domrong et al., 2022) . . . . .	50

A.5	The top 50 collocations from Korean Wikipedia. (Cheevaprawatdomr- rong et al., 2022) . . . . .	51
A.6	The top 50 collocations from Thai Wikipedia. (Cheevaprawatdomrong et al., 2022) . . . . .	52
A.7	The top 50 collocations from Arabic Wikipedia. (Cheevaprawatdom- rong et al., 2022) . . . . .	53
B.1	The topic keys from the New York Times with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	55
B.2	The topic keys from the United States State of the Union Addresses with different retokenization measures. (Cheevaprawatdomrong et al., 2022) .	56
B.3	The topic keys from the Yelp Dataset with different retokenization mea- sures. (Cheevaprawatdomrong et al., 2022) . . . . .	57
B.4	The topic keys from the the Ten Thousand German News Articles Dataset with different retokenization measures. (Cheevaprawatdom- rong et al., 2022) . . . . .	58
B.5	The topic keys from the Chinanews with different retokenization mea- sures. (Cheevaprawatdomrong et al., 2022) . . . . .	59
B.6	The topic keys from the Dianping with different retokenization mea- sures. (Cheevaprawatdomrong et al., 2022) . . . . .	60
B.7	The topic keys from the Douban with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	61
B.8	The topic keys from the Webhose’s Free Datasets with different retok- enization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	62
B.9	The topic keys from the KAIST Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	63
B.10	The topic keys from the Prachathai with different retokenization mea- sures. (Cheevaprawatdomrong et al., 2022) . . . . .	64
B.11	The topic keys from the Wongnai with different retokenization mea- sures. (Cheevaprawatdomrong et al., 2022) . . . . .	65

B.12 The topic keys from the BEST Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	66
B.13 The topic keys from the Thai National Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	67
B.14 The topic keys from the Antcorpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022) . . . . .	68



# Chapter I

## INTRODUCTION

### 1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) model is an unsupervised statistical model that allows us to find topics or themes in a large set of text documents (Blei et al., 2003). The LDA receives words as input and produces topics, which are probability distributions of words in the corpus. A topic is often represented by a list of high probability words within the topic called topic keys. The documents are expressed as combinations of such topics.

### 1.2 Topics

Figure 1.1 contains some examples of topics from restaurant reviews, represented by their topic keys. There are topics about desserts, burgers, Asian foods, beverages, and breakfast. In topic 1, since the topic is about desserts, we can guess that the words “ice” and “cream” are actually part of the word “ice cream,” which conveys more meaning than the individual words. The topic keys could be more meaningful if it presents “ice cream” together, as in the following example.

chocolate, cake, **ice cream**, dessert, try, sweet, love, make, one

**Topic 1:** chocolate, cake, cream, dessert, try, sweet, love, make, ice, one

**Topic 2:** burger, fry, burgers, order, get, hot, good, cheese, dog, like

**Topic 3:** food, chicken, dish, thai, order, rice, chinese, soup, good, pho

**Topic 4:** beer, bar, good, great, beers, food, place, drink, selection, wing

**Topic 5:** breakfast, egg, order, good, food, bacon, wait, toast, place, pancakes

Figure 1.1: Example of topics from restaurant reviews.

The same can be said about topic 2, where the topic could be more meaningful if the words “hot” and “dog” are together as “hot dog.” When “hot” and “dog” are separated, readers may think that “dog” refers to an animal. The meaning of the word “dog” will be inconsistent with other words in the topic about burgers, resulting in an incoherent topic.

### 1.3 Collocations

The groups of words, such as “ice cream” and “hot dog,” are called collocation, which is two or more words that convey conventional meaning. The meaning of the collocations often goes beyond the meaning of the components. For example, the word “super bowl” refers to a football match, which is the meaning not captured by the words “super” and “bowl.”

The topic would be more meaningful and more coherent when collocations are presented together, but in the traditional LDA they are not as we feed the model with individual words. In addition, the model is based on the bag-of-words assumption, which assumes that the order of the words does not matter. As a result, the original meaning of the words, such as “ice cream,” “hot dog,” and “super bowl,” could be lost.

To fix this problem, we can group together some input words with a significant relationship, and feed them to the LDA model as a single unit to preserve their special meaning. The question then comes to how we decide which group of words should be connected. However, before discussing how to connect the words, we first need to extract them from our documents.

### 1.4 Tokenization

Typically, the data in the document is in the form of text, sentence after sentence, which is not the form of input the LDA needs. In order to obtain input words

for the LDA model, we perform tokenization, which is a process of breaking text into chunks of words called tokens. Although breaking text may seem obvious for a language with explicit word boundaries such as English, it is not that simple. Periods, which often suggest the end of words, could be a part of an abbreviation. Hyphens, which are often within a word, such as “e-mail” and “co-occur”, could also separate the two words, such as in “San Francisco-Los Angeles flights”. Collocations may also be a problem where the whitespace between two words may not suggest they should be separated.

Tokenizing issues are even more crucial in the languages that do not have clear word boundaries, including Chinese and Thai. For example, we need a tokenizer to break a sentence “นายกรัฐมนตรีเดินทางไปต่างประเทศ (the prime minister travels abroad)” into words “นายก [na jok] (prime), รัฐมนตรี [rad t<sup>h</sup>a mon tri ] (minister), เดินทาง [d n t<sup>h</sup>an] (travel), ไป [paj] (to), ต่าง [ta ŋ] (other), ประเทศ [pra t<sup>h</sup>e d] (country).”

Although there are some tokenizers that work well in these languages, there is no single tokenizing standard that works well for all tasks. A good tokenizer is the one that does the right amount of breaking. If it breaks the text too much into smaller tokens, we may end up losing the original meaning of the words. On the other hand, if the tokenizer doesn't break enough, we could see many unnecessary longer distinct words, which expand the vocabulary size of the model or increase out-of-vocab problems.

Modern tokenizers are often built using machine learning techniques. They are trained on annotated data, where linguists mark the word boundaries in the training document. Therefore, the criteria linguists used to segment text into tokens affect the standard of the tokenizer. When the annotators prefer to break text into smaller words, the resulting tokenizer tends to break collocations into separated individual words, which leads to a meaning loss in downstream tasks.



## 1.5 Retokenization

A remedy to the problem of lost meaning due to the tokenization, as we mentioned, is to do the retokenization, or merging the input words with special relationships after the tokenization process. Many strategies can be employed to help decide whether each pair of adjacent words should be merged. In this research, we explore three such merging strategies, including chi-squared statistics,  $t$ -statistics, and raw frequency counts of phrases. We believe that many languages, especially the ones that do not have clear tokenization standards, deserve investigation into what kind of processing is appropriate.

## 1.6 Objective and Hypothesis

The objective of this research is to study the performance of different merging measures when they are employed in different types of languages. In particular, we are interested in the strategies that would make the resulting topics generated by the LDA model more coherent and meaningful, as well as increase the goodness of fit of the model, or how well the model represents the data.

We hypothesize that the strategies could influence the goodness of fit of the model and the coherence of the topics. We perform experiments on English, German, Chinese, Japanese, Korean, Thai, and Arabic, a set of languages with different writing systems and morphological typology, to understand how the merging strategies perform in various types of language.

## 1.7 Contribution

The main contributions of this research are as follows:

- The results of the experiment show that a  $t$ -statistic and raw-frequency merging measures can improve the results of the LDA across all language types

and writing systems when the input documents do not differ much from the collocation training data.

- The study found that the results tend to be better when more tokens are merged.
- The investigation indicates that when we use  $\chi^2$  measure to produce a truncated list of collocations, the resulting list rarely merges any collocation in input documents. Therefore, this strategy is less suitable for merging the input of topic models.



## Chapter II

# BACKGROUND AND LITERATURE REVIEW

## 2.1 Latent Dirichlet Allocation

### Model

Latent Dirichlet Allocation (LDA) is a topic model used to find hidden topics from unlabeled and unannotated documents. We say the topics are hidden because we can only see words but not the underlying topics when we observe the documents. Figure 2.1 shows the intuition of the model. Discovering these topics provides us with useful insight into what different themes there are in the documents. Since the model is unsupervised, it is handy in processing a large number of documents.

A topic consists of words with their probabilities. We often represent a topic by its topic-keys which are the top words in the topic. For example, in a topic about baseball, topic-keys would include “san, francisco, baseball, yankees, game, league, last, first, chicago, run, mets, season, stadium, team, pitch, diego, today, manager, major, home.” In Figure 2.2, we can find themes or topics about education, sport, military, police, healthcare, Europe, president, court, music, and restaurant from the New York Times corpus.

There are a few assumptions of the model.

1. We assume that the order of the words in the document does not make a difference. Therefore, we can use a bag of words as input.
2. We assume that the order of the document does not make a difference.
3. Although we do not know what each topic looks like, we assume that we know the number of topics in the documents.

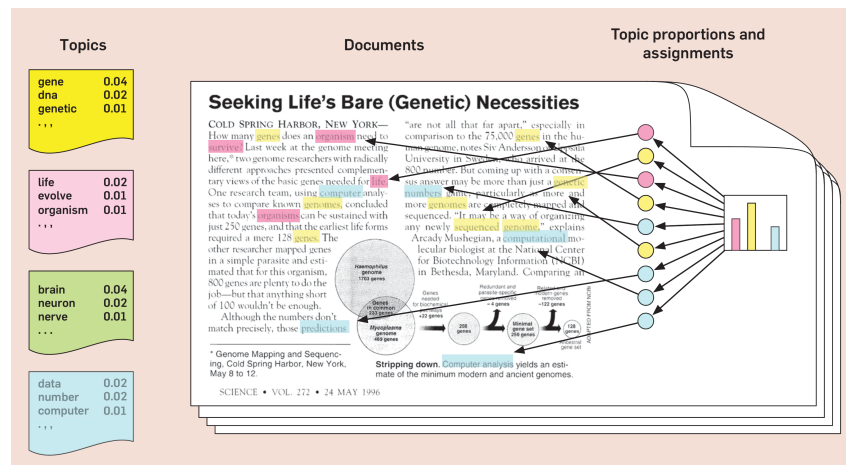


Figure 2.1: The intuition of the LDA model. (Blei, 2012)

- Topic 1:** school, university, high, students, college, study, new, education, public, graduate  
**Topic 2:** team, national, coach, football, game, season, league, basketball, players, play  
**Topic 3:** today, say, united, military, war, american, states, force, army, officials  
**Topic 4:** say, police, kill, people, fire, man, two, officer, shoot, yesterday  
**Topic 5:** health, drug, say, job, care, people, use, make, workers, work  
**Topic 6:** london, world, european, german, war, europe, today, west, germany, british  
**Topic 7:** washington, president, today, reagan, administration, house, bush, say, clinton, white  
**Topic 8:** court, judge, federal, rule, state, right, supreme, appeal, say, case  
**Topic 9:** music, dance, hall, night, concert, new, program, opera, work, theater  
**Topic 10:** food, restaurant, eat, wine, restaurants, cook, use, drink, fresh, sell

Figure 2.2: Example of topics from New York Times corpus.

## Generative Process

The objective is to find hidden topics from the words in the documents we observe. However, instead of going directly to the process of discovering the topics, we first look at the generative process, or how we can generate a document given that we know the hidden topics and the hidden topic distribution.

1. Choose a topic distribution
2. For each word,
  - a) Randomly choose a topic from the topic distribution
  - b) Randomly choose a word from the topic, which is a distribution of words.

The joint distribution of the hidden and observed variables are as follows.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right) \quad (2.1)$$

where  $\beta_k$  is topic  $k$  which is a distribution of words,  $\theta_d$  is the topic proportion of the  $d$ th document,  $z_d$  is the topic assignment of the  $d$ th document,  $z_{d,n}$  is the topic assignments for  $n$ th word in the  $d$ th document,  $w_d$  is the observed words in  $d$ th document, and  $w_{d,n}$  is the  $n$ th word in the  $d$ th document.

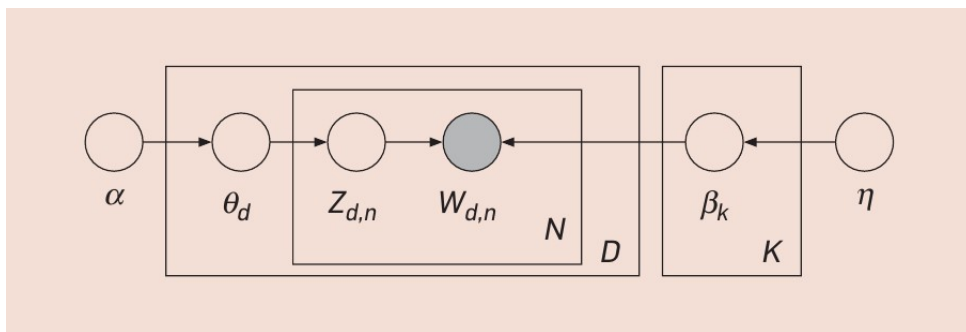


Figure 2.3: The figure shows the relationship of random variables (Blei, 2012)

## Topics Discovery Process

The process of discovering topics, which is what we want, can be understood as a reversion of the generative process. We try to compute the posterior distribution given that we have observed the words in the documents.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D})}{p(w_{1:D})} \quad (2.2)$$

We can compute  $p(\beta_{1:K}, \theta_{1:D}, z_{1:D})$ , which is the joint distribution of the random variables. However, it is difficult to compute the  $p(w_{1:D})$ , which is the probability of seeing this corpus from any possible combination of the hidden topics because such a number of combinations is huge. Statisticians use algorithms to approximate this posterior distribution. One of the widely used algorithms is a sampling algorithm called Gibbs sampling.

## Gibbs Sampling

For the words  $\mathbf{w} = \{w_1, \dots, w_n\}$ , where each word  $w_i$  is in document  $d_i$ , our objective is to discover the  $\theta_{d_i}$ , which is the distribution of topics in document  $d_i$  and the  $\beta_j$ , which is the distribution of the words in topic  $j$ . We would be able to estimate both  $\theta_{d_i}$  and  $\beta_j$  if we have all  $z_i$ , which is the topic assignment for word  $w_i$ .

Gibbs sampling uses the Markov Chain Monte Carlo (Gilks et al., 1995), which is an algorithm for sampling from a probability distribution. Each step of the Gibbs sampling try to assign the topic  $z_i$  for each word  $w_i$  in the document by sampling from the probability distribution calculated from equation 2.3 (Griffiths and Steyvers, 2004).

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \gamma}{n_{-i,j}^{(\cdot)} + W\gamma} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \quad (2.3)$$

where  $\mathbf{z}_{-i}$  is the topic assignment for the words  $w_k, k \neq i$ ,  $n_{-i,j}^{(w_i)}$  is the number of word  $w$  assigned to topic  $j$ , not including this current word,  $n_{-i,j}^{(\cdot)}$  is the total number of words assigned to topic  $j$ , not including this current word,  $\gamma$  is a smoothing parameter,  $W$  is the number of words in the vocabulary,  $n_{-i,j}^{(d_i)} + \alpha$  is the number of words from document  $d_i$  that is assigned to topic  $j$ , not including this current word,  $n_{-i,\cdot}^{(d_i)}$  is the total number of words in document  $d_i$ , not including this current word,  $T$  is the number of topic, and  $\alpha$  is a smoothing parameter.

This probability in the equation gives a chance that topic  $j$  will be assigned to the word  $w_i$  which is depending on the first fraction, which can be interpreted as the smooth version of the current proportion of the word  $w_i$  assigned to topic  $j$ , and the second fraction, which can be explained as the smooth version of the proportion of the word  $w_i$  assigned to topic  $j$  in this document  $d_j$ .

During the iterations, the algorithm will keep adjusting the topic assignment of each word. Eventually, the assignments are going to be good enough to be used to calculate the  $\theta_{d_i}$  and  $\beta_j$  that we are interested in.

Mallet (McCallum, 2002), which is the toolkit we use to discover the hidden

topics in this research, is a fast and highly scalable implementation of the Gibbs sampling.

## 2.2 Improving and Evaluating Topic Models

### Preprocessing for Topic Models

Many studies show that the results of LDA depend on preprocessing steps even in a language with word boundaries such as English

May et al. (2016) studies the effect of lemmatization on the interpretability of topic models in Russian, a language with high morphological variation. They found that interpretability improves when the corpus contains untruncated documents, the vocabulary is filtered, and lemmatization is used.

Schofield et al. (2017) found that removing stopwords improve the coherence of topic models. However, it is sufficient to remove the most common, evident stopwords from a corpus without constructing a specific stoplist for the problem.

### Topic Models with Phrases-based Input

Many works recognize that LDA results can be improved when input includes phrases

Lindsey et al. (2012) present an extension to Latent Dirichlet Allocation called Phrase-Discovering Latent Dirichlet Allocation. The model uses a hierarchy of Pitman-Yor processes to infer the location, duration, and topic of phrases within a corpus while relaxing the bag-of-words assumption. The experiment on human subjects shows that the algorithm finds significantly better, more interpretable relevant phrases than competing models.

Lau et al. (2013) shows that including bigram collocations in the document representation leads to better topic coherence. The research finds that a small num-

ber of top-ranked bigrams, up to 1000, improves subject quality when compared to unigram tokenization. Using up to 10,000 bigrams can improve topic quality even further. The paper also shows that named entities with several words provide consistent results, implying that they should be represented as single tokens.

Yu et al. (2013) provides a phrase-based LDA model that uses a key phrase extraction methodology, the C-value method, to transition from a bag of words or n-grams paradigm to a "bag-of-key-phrases" to discover latent themes. The paper demonstrates that the model can help LDA create better and more interpretable themes than those generated using the bag-of-n-grams technique by employing a phrase incursion user study.

El-Kishky et al. (2014) presents ToPMine, a topical phrase mining framework for segmenting documents into single and multi-word phrases, and a new topic model that works with the induced document partition.

Wang et al. (2016) presents PTR, which is a phrase-based topical ranking algorithm for extracting key words from scientific papers. Candidate keys are separated into different themes and used as vertices in a topic's phrase-based graph. Then, to rank phrases for each topic, PageRank is partitioned into several weighted-PageRanks. Keyphrases are finally identified by their overall scores on related connected topics.

Bin et al. (2018) presents a key phrase and LDA model-based strategy for discovering and recommending hot subtopics in Chinese news. The study chooses the Longest Common Sequence (LCS) value as the similarity distance during the clustering of hot subtopics.

Li et al. (2018) combines a quality phrase mining method and a document clustering method to provide topical cohesion.



## Evaluating Topic Models

Evaluating the results of LDA can be complicated. One must assess the statistical fit of the model, as well as the coherence of the topic keys, which are the most probable words in each topic. However, the two measurements may not agree. As Chang et al. (2009) points out, topic models that outperform on held-out likelihood may infer less semantically meaningful topics. Researchers usually use the evaluations of fit (Wallach et al., 2009), together with the measure of coherence based on mutual information (Bouma, 2009; Mimno et al., 2011).

The analyses often require the models in focus to have the same vocabulary and tokenization standard. These requirements do not hold for our study. Schofield and Mimno (2016) proposes a metric called the normalized log-likelihood per token, which measures how much the log-likelihood per token of the model in focus has improved over that of the original model without requiring the two models to have the same number of vocabulary.

### 2.3 Collocations

Collocations consist of two or more words that express conventional meaning (Manning and Schutze, 1999). Examples of collocations include noun phrases, such as “New York,” which refers to a city in the United States. It conveys conventional meaning beyond the content of its two components, “New” and “York.” Collocations also consist of phrasal verbs such as “knock down” and “build in.” These groups of words co-occur so frequently that native speakers use them correctly. For instance, we use “strong tea” and “powerful car” but not “powerful tea” or “strong car.”

Understanding collocations is useful in a variety of situations. It ensures that the final sentences sound genuine and without errors when text is automatically generated. It can also automatically determine which collocations should be included in a dictionary entry. In addition, it can help improve the parsing of text that includes

collocations.

There are many strategies we can use to identify the collocations, including frequency, mean and variance, and hypothesis testing.

## Frequency

Frequency is the simplest way to find collocations. If two words often appear together, they could have a special relationship and, therefore, could be a collocation.

Ranking phrases by frequency crudely may not yield interesting results because the top frequent pairs may include collocations without much meaning, such as “of the,” “in the,” and “to the.” One way we can filter these uninteresting phrases is by using a part-of-speech filter to look for pairs that could be “phrases.” We can also eliminate so-called stopwords or words that do not add much meaning before counting the frequency.

Frequency can also be used to suggest a correct phrase. In table 2.1, we can clearly observe that “strong support” and “powerful symbol” are correct while “powerful support” and “strong symbol” are not.

## Mean and Variance

The words in a collocation may not always be next to each other. The distance between them can be flexible. For example, as we can see in these two sentences:

- The girl takes care of her ailing cat.
- She always takes good care of other people around her.

The distance between the word “takes” and “care” can vary. Therefore, we can understand the relationship between the two words we are interested in by computing

w	C(strong, w )	w	C(powerful, w )
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	wepons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	9	magnet	4

Table 2.1: The nouns w occurring most often in the patterns “strong w” and “powerful w” from New York Times newswire (Manning and Schutze, 1999).

the mean and the variance of their distance. For example, the distance between the word “takes” and “care” in the samples are 1 and 2.

The mean ( $\bar{d}$ ) is defined as

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad (2.4)$$

where  $d_i$  are the distances between two words, and  $n$  is the number of the pair. And the variance  $s^2$  is defined as

$$s^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1} \quad (2.5)$$

The low variance of the distance between two words indicates an interesting relationship. High variance means that the relationship is quite random and, therefore, not so interesting. In Figure 2.4, the lower variance of the distance between two words suggests that “strong opposition” and “strong support” are more interesting collocations than “strong for.”

## Hypothesis Testing

Hypothesis testing is a technique for drawing statistical conclusions from population data. It is used to determine whether or not the outcomes of an experiment are meaningful. It starts with declaring a null hypothesis, which is a statement that says there is no significant difference in the given observations, and an alternative hypothesis, which is a contradiction statement of the null hypothesis. Then statisticians collect data and analyze the data to either reject the null hypothesis or conclude that the observed difference is insignificant and only happens by chance.

In our work, hypothesis testing is used to determine whether a pair of words co-occur by chance. The null hypothesis  $H_0$  is formulated as the two words co-occur only by chance, and there is no significant relationship between them. The alternative hypothesis  $H_1$  would be the opposite of  $H_0$ , which is that the two words

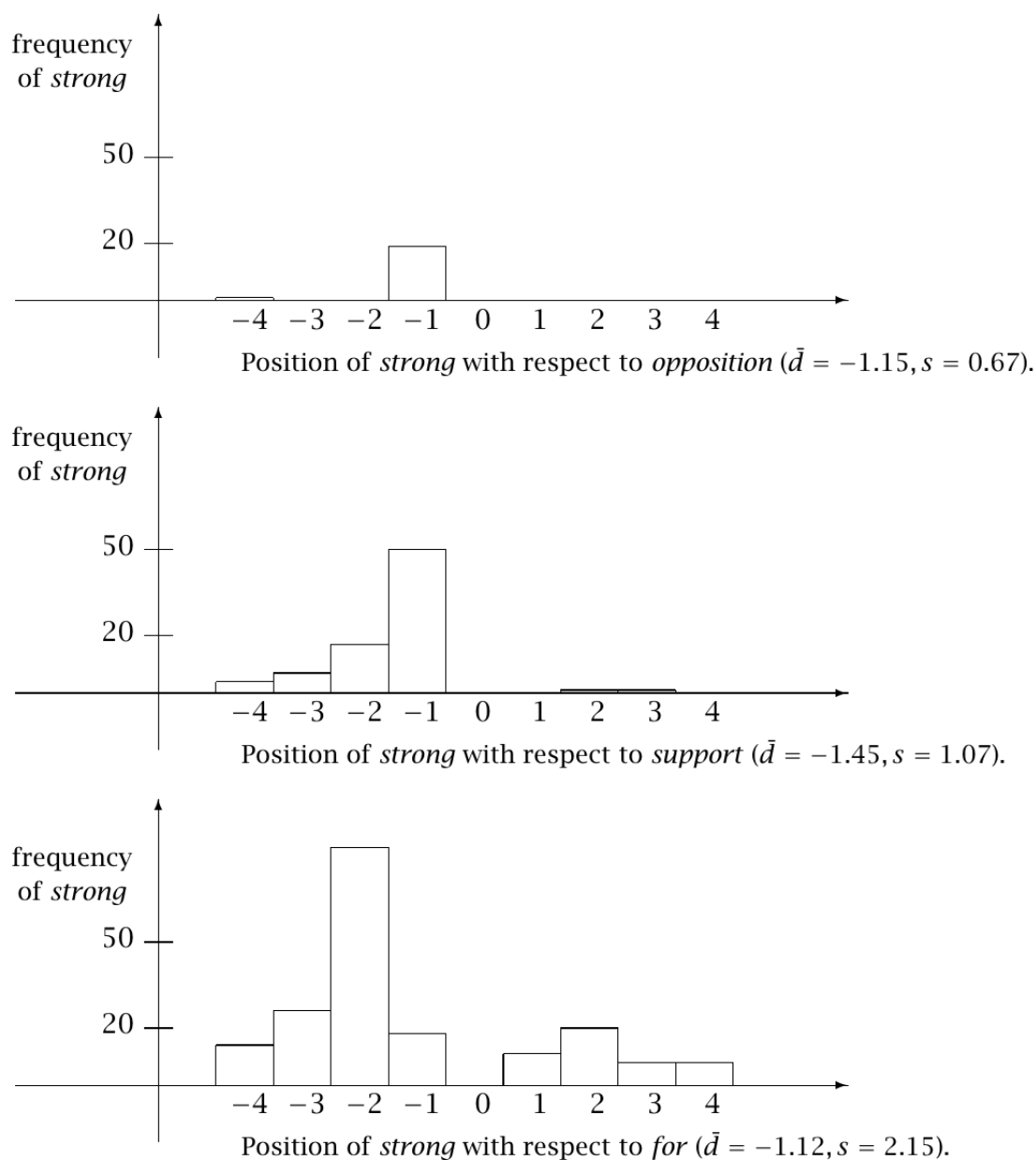


Figure 2.4: Histogram of the position of strong relative to three words from the New York Times newswire (Manning and Schutze, 1999)

are somehow associated. Then we assume  $H_0$  is true and compute  $p$ , the probability that the two words would appear as they are. If the probability  $p$  is too small, which means if  $H_0$  is true, it is unlikely that the observation would be as we have seen, we reject the null hypothesis  $H_0$ . Otherwise, we retain  $H_0$ .

### The $t$ test

We can apply the  $t$  test (Student, 1908) to decide whether to reject the null hypothesis. The  $t$  statistic is defined as

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \quad (2.6)$$

where  $\bar{x}$  is the sample mean,  $s^2$  is the sample variance,  $N$  is the sample size,  $\mu$  is the distribution mean. After we compute the  $t$  we reject the null hypothesis if  $t$  value is larger than the critical value for our preferred confidence level.

In our collocation problem, we first assume that the null hypothesis is true, or in other words,  $w_1$  and  $w_2$  co-occur only by chance. Then, we generate bigrams randomly. When the bigram is  $w_1$  and  $w_2$ , we says that the outcome is 1, and when the bigrams is not  $w_1$  and  $w_2$ , we says that the outcome is 0. This is a Bernoulli process with  $p = P(w_1)P(w_2)$ . where  $P(w_1)$  and  $P(w_2)$  are the probabilities that the word  $w_1$  and  $w_2$  occur respectively.

The sample mean ( $\mu$ ) for this process is  $P(w_1, w_2)$ , and the sample variance ( $s^2$ ) is  $P(w_1, w_2)(1 - P(w_1, w_2))$ , where  $P(w_1, w_2)$  is the probability that two adjacent words are  $w_1$  and  $w_2$ . this sample variance is approximately  $P(w_1, w_2)$  when  $P(w_1, w_2)$  is very small.

The equation 2.6 then becomes

$$t(w_1, w_2) \approx \frac{P(w_1, w_2) - P(w_1)P(w_2)}{\sqrt{\frac{P(w_1, w_2)}{N}}} \quad (2.7)$$

If, for example, in a corpus consisting of 15245658 words, we found the word “hot issues” 7 times, “hot” 15256 times, and ”issues” 4258 times. We can calculate

	$w_1 = \text{some}$	$w_1 \neq \text{some}$
$w_2 = \text{definition}$	10	5786
$w_2 \neq \text{definition}$	14325	15668574

Table 2.2: The number of occurrences of the word “some definition” and other words

the  $t$  statistic as follows.

$$t \approx \frac{7}{15245658} - \frac{15256}{15245658} \times \frac{4258}{15245658} \quad (2.8)$$

$$\approx \frac{7}{15245658} - \frac{15256 \times 4258}{15245658^2} \quad (2.9)$$

Since the critical value for  $\alpha = 0.005$  is 2.576 we cannot reject the null hypothesis that “hot issues” co-occur only by chance, and there is no significant relationship between the two words “hot” and “issues.”

### Pearson’s chi-square test

Pearson’s chi-square test (Pearson, 1900) is an alternative test. It is different than the  $t$  test because it does not assume that the probability are approximately normally distributed. The chi-square ( $\chi^2$ ) statistic is defined as

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2.10)$$

where  $i$  and  $j$  are the row and column in the table, while  $O_{ij}$  and  $E_{ij}$  are the observed and expected value for cell  $(i, j)$ . For the case of 2 by 2 table, in which column 1 is for  $w_1$ , column 2 is for not  $w_1$ , row 1 is for  $w_2$  and row 2 is for not  $w_2$ , we can work out the Equation 2.10 to get

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \quad (2.11)$$

Again, we reject the null hypothesis if  $\chi^2$  exceeds the critical value for our significance level.

Suppose we observe the number of occurrences in a corpus as in Table 2.2, we can compute the  $\chi^2$  statistic as follows.

$$\chi^2 = \frac{15688695(10 \times 15668574 - 5786 \times 14325)^2}{(10 + 5786)(8 + 15820)(5786 + 15668574)(14325 + 15668574)} \quad (2.12)$$

$$\approx 3.79 \quad (2.13)$$

Since the critical value is 3.841 for  $\alpha = 0.05$  we cannot reject the null hypothesis that “some definition” co-occur only by chance, and there is no significant relationship between the two words “some” and “definition.”

The chi-square ( $\chi^2$ ) statistic can also be computed by

$$\chi^2(w_1, w_2) = \frac{N(P(w_1, w_2) - P(w_1)P(w_2))^2}{P(w_1)P(w_2)} \quad (2.14)$$

## Likelihood Ratios

Hypothesis testing can also be done using likelihood ratios, especially when data are sparse. The likelihood ratio is relatively more interpretable than the  $\chi^2$  statistic. It describes the ratio of the chance one hypothesis has over the other (Dunning, 1993).

We can formulate two hypotheses when we are interested in a bigram  $w_1 w_2$ . The first hypothesis states that  $w_1$  and  $w_2$  are independent.

$$\text{Hypothesis 1: } P(w_2 | w_1) = p = P(w_2 | \neg w_1)$$

The second hypothesis states the opposite, saying that they are dependent, which means they can be an interesting collocation.

$$\text{Hypothesis 2: } P(w_2 | w_1) = p_1 \neq p_2 = P(w_2 | \neg w_1)$$



We calculate  $p, p_1, p_2$  as follows.

$$p = \frac{c_2}{N} \quad (2.15)$$

$$p_1 = \frac{c_{12}}{c_1} \quad (2.16)$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1} \quad (2.17)$$

where  $c_1$  is the number of times  $w_1$  occurs,  $c_2$  is the number of times  $w_2$  occurs, and  $c_{12}$  is the number of times  $w_1 w_2$  occurs.

Assuming a binomial distribution we can compute the probability as follows.

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \quad (2.18)$$

The likelihood for hypotheses 1 and 2 would be

$$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p) \quad (2.19)$$

$$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2) \quad (2.20)$$

then the log of the likelihood ratio  $\lambda$  would be

$$\begin{aligned} \log \lambda = & \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\ & - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2) \end{aligned} \quad (2.21)$$

where  $L(k, n, x) = x^k (1-x)^{n-k}$ .

For example, the word “powerful force” occurs 10 times in the New York Times corpus, while the individual word “powerful” and “force” occur 932 and 3424 times respectively. We can calculate  $-2 \log \lambda = 80.39$ , which can be interpreted that the two words “powerful force” are  $e^{0.5 \times 80.39} = 2.86 \times 10^{17}$  times more likely than the would by random chance.

And since  $-2 \log \lambda$  is asymptotically  $\chi^2$  distributed (Mood et al., 1974), we can also utilize the number  $-2 \log \lambda = 80.39$  to reject  $H_1$  for  $\alpha = 0.005$  since 82.96 is more than the critical value of 7.88.

## 2.4 Morphological Typology and Writing Systems

### Morphology

In linguistics, morphology is a study of the structure within words. It focuses on how words are constructed from smaller meaningful components and how such construction affects the meaning or grammatical function of the final word (Dawson et al., 2016). Morphological processes include derivation and inflection.

#### Derivation

Derivation is the process of constructing words out of other words. Derivation takes a single word and applies one or more operations to it, yielding a new term, generally belonging to a different lexical category, or sometimes called part of speech (Dawson et al., 2016). An example of derivation is constructing the word “co-author” from the word “author” by adding “co-” in front of the original word to create a new word that means joint author.

#### Inflection

Inflection is the process of changing the grammatical forms of words. Inflection uses stems and affixes or other processes as derivation does. However, the critical difference is that, instead of developing wholly new words, it changes forms of the original words (Dawson et al., 2016). An example of inflection would be to create the word “dogs” from “dog” by adding “-s” to develop a plural version of the original word.

### Morphological Typology

Morphological typology is a study that tries to classify languages by grouping them based on their morphological patterns. Languages are categorized by whether or not they employ morphological processes in analytic or synthetic languages.

## Analytic Languages

Analytic languages are languages that are constructed from sequences of free morphemes. Each word is made up of a single morpheme with meaning and function. Separate words are used in analytic languages to represent semantic and grammatical notions that are commonly expressed with affixes in other languages (Dawson et al., 2016). Chinese is an example of an analytic language. In Chinese, the concepts of plurality and the past tense are expressed through the use of function words, not a change in form. As shown in Figure 2.5, the word [le] is used to show past tense.



Figure 2.5: Chinese is an example of an analytic language

## Synthetic Languages

In synthetic languages, a word can be constructed by connecting many meaningful morphemes. The added morpheme can communicate grammatical functions or provide additional meaning to the word. There are many types of synthetic languages, including agglutinating languages and fusional languages.

### **Agglutinating Languages**

The morphemes in agglutinating languages are linked together loosely. Therefore, it is typically simple to discover where the morpheme boundaries are. For example, in Hungarian words [ha z-unk] (our house) and [ha z- d] (your house) we can see that [unk] means “our” and [ d] means “your” (Dawson et al., 2016)

### **Fusional Languages**

Words in fusional languages are generated by adding bound morphemes to stems, like in agglutinating languages. However, the affixes may be difficult to be separated from the stem because of the fusion between morphemes. For example, in Spanish, “hablo,” “habla,” and “hable” means “I am speaking,” “S/he is speaking,” and “I spoke” respectively. But we cannot conclude that “habl” means speak because there is no such a morpheme in Spanish. Actually, “hablar” would mean speak. We can observe the fusion of going from “hablar” to “hablo,” “habla,” and “hable.”

In agglutinating languages, each affix typically indicates only one meaning, whereas, in fusional languages, a single prefix commonly transmits multiple meanings simultaneously.

### **Influence of Morphological Typology on Latent Dirichlet Allocation**

Since Latent Dirichlet Allocation uses words as input, the text needs to be broken into tokens by using tokenizers. Each language has their complication when we try to break long sentences into tokens. For example, in German, compound nouns are written without spaces, so tokenizers have to perform the task of breaking these compound nouns. However, in the process of breaking compound nouns, together with the bag-of-word assumption of the model, the meaning of broken compound nouns could be lost. Therefore, the morphological characteristics of each language may affect the suitability of retokenization strategies that try to merge the words after the tokenization processes.

## Writing Systems

Writing is the use of graphic marks to represent specific linguistic utterances (Rogers, 2005). Writing helps us to communicate with other people beyond using spoken language. As we write, we record a specific thought for a specific audience.

The components of writing are called graphs or graphemes. We can categorize a writing system based on whether the graphemes of that system are primarily used to express sound or meaning. Phonographic systems, such as English, base primarily on the representation of sound, while morphographic systems, such as Chinese, rely on the representation of meaning.

### Logograms

Logograms rely on a relationship between a written graphemes and a specific word or morpheme, mainly its meaning. The symbols may or may not give information about the pronunciation. An example of logograms are Chinese and Sumerian cuneiforms. In Figure 2.6, we can see that each of the chinese character denotes a specific meaning.

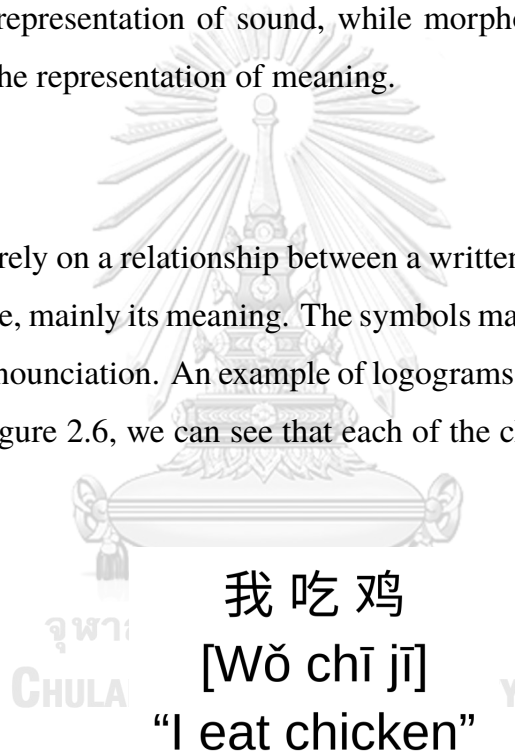


Figure 2.6: Chinese is an example of logogram

### Syllabary

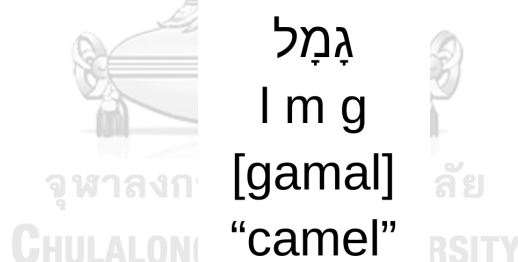
In a syllabary, such as Japanese Hiragana, characters denote syllables, In Figure 2.7, we can see that each character represents a syllable.

わたしはにほんじんです  
 [wa ta shi wa ni hon jin de su]  
 “I am Japanese”

Figure 2.7: Japanese is an example of syllabary

### Abjads

Abjads are systems that represent consonants but not vowels. Some examples of abjads are Arabic and Hebrew. Reading without vowels may seem complicated, but language understanding usually allows the reader to add the vowels by looking at the context of a sentence. In Figure 2.8, the first line presents Hebrew characters, and the second line shows the corresponding consonants in English. We can see that there is no character denoting vowels. The readers must fill in vowels to read it as [gamal], which means camel in English.



גמל  
 l m g  
 [gamal]  
 “camel”

Figure 2.8: Hebrew is an example of abjad

### Alphabet

In alphabet writing systems like English and German, characters denote both consonants and vowels. For example, in a word “hit,” “h” and “t” denote consonants, and “i” denotes vowels.

## Abugidas

Abugidas are writing systems that use characters to denote consonants and specific characters to denote vowels. Thai is an example of Abugida. In Figure 2.9, “ดีใจ” is constructed by using a sequences of “ด” (consonant) “ี” (vowel) “จ” (consonant) “ใจ” (vowel) “จ” (consonant)

ดีใจ  
[di: chaj]  
“happy”

Figure 2.9: Thai is an example of abugida

## Featural

Another writing system is featural, where character denotes place and manner of articulation, together with voicing, of phonemes. Korean is an example of a featural script. In Figures 2.10, the word [Hangul], which means “Korean,” has two syllables and is denoted by two characters. In the first character, there are three components. The upper left component denotes the [h], the upper-right component represents [a], and the lower component indicates [n]. In the second character, the upper line denotes [g], the longest horizontal line in the middle represents [eu], and the lower component indicates the ending [l].

한글  
[Hangul]  
“Korean”

Figure 2.10: Korean is an example of featural system

## **Influence of Writing Systems on Latent Dirichlet Allocation**

Tokenization is an important preprocessing step that prepares token input for Latent Dirichlet Allocation model. However, the resulting tokens are not in the same standard because the tokenization processes differ between languages, part of that because they use different writing systems. Therefore each retokenization strategy may produce different results when they try to merge tokens with different characteristics from many writing systems.





## Chapter III

### OUR PROPOSED METHOD

#### 3.1 Collocations as LDA Token

We hypothesize that merging words into ngrams would increase the coherence of the topic keys. The merging can also be helpful, particularly in languages without clear word boundaries. It can help adjust the tokenizing standard to be more appropriate for topic modeling.

There are many ways to merge words into collocations (Manning and Schütze, 1999). We analyze the use of chi-squared statistics ( $\chi^2$ ), the t-statistic, and raw frequency to compute the threshold to decide whether or not adjacent words are merged.

We first compute the collocation measures in a large corpus. Then for each measure, we list the top 50,000 bigrams with the highest scores. We use this list to merge the input tokens of the LDA.

#### 3.2 Evaluation Metrics

To study the contribution of merging input tokens to the result of the LDA, we measure the improvement of statistical fit and coherence with held-out likelihood and a silhouette coefficient based metric.

##### Held-Out Likelihood

Combining words into phrases gives us fewer tokens and a larger vocabulary size. Therefore, we cannot use the conventional log-likelihood metric to compare the word-token and collocation-token models. To account for the difference in the

number of tokens and the vocabulary size, we normalize the log-likelihood by dividing it with the log-likelihood of the null (unigram) model as in Schofield and Mimno (2016). Then, the normalized log-likelihood per token ( $\text{PTLL}_{\text{norm}}$ ) is

$$\text{PTLL}_{\text{norm}} = \frac{\log \mathcal{L}_{\text{model}} - \log \mathcal{L}_{\text{unigram}}}{N} \quad (3.1)$$

where  $N$  is the number of tokens. This metric measures how the log-likelihood per token of the collocation-token model has improved over the log-likelihood per token of the word-token model. Since this metric has already been normalized, the model with higher PTLL is better.

## Concatenation-based Embedding Silhouette (CBES)

Metrics used to measure topic coherence assume that all models have the same vocabulary. This is not our case. Therefore, we would like to propose the new application of the silhouette coefficient (Rousseeuw, 1987), which is a standard metric for evaluating clustering.

In a good topic, topic keys should be close to each other and away from the topic keys in other topics. In our case, the length between topic keys is measured by the cosine distance between word embedding. So, the cosine distance should be relatively smaller within the topic and relatively larger between the keys from different topics.

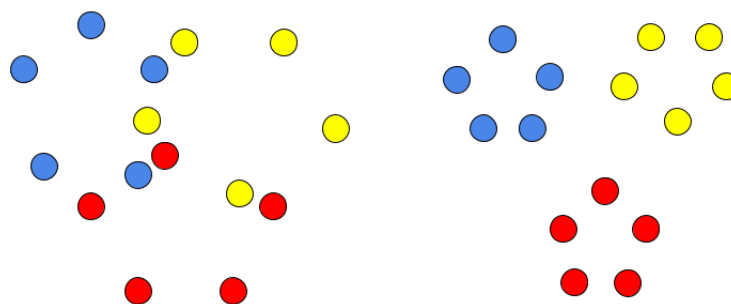


Figure 3.1: The topic on the right is a better topic

To compute the silhouette coefficient, we first calculate the  $a(i)$ , which is the mean cosine distance between topic-key  $i$  and other topic-keys in the same topic.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \quad (3.2)$$

where  $d(i, j)$  is the distance between  $i$ th and  $j$ th topic-key. Then for each other topic, we calculate the mean of the distance of topic-key  $i$  to topic-keys in that other topic. And  $b(i)$  is the smallest of such mean among other topics.

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \quad (3.3)$$

After we get  $a(i)$  and  $b(i)$ , the silhouette coefficient for topic-key  $i$  is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \text{ if } |C_i| > 1 \quad (3.4)$$

and

$$s(i) = 0, \text{ if } |C_i| = 1 \quad (3.5)$$

The silhouette coefficient for the entire model is the average  $s(i)$  over all  $i$ . The larger silhouette coefficient means that topic-keys are relatively similar within its topic and different from other topics.

### 3.3 Experiments

We think that we should consider the morphology of the language when we choose the pre-processing steps. We study the corpus from different morphological typologies, such as German, which is a fusional language; Japanese and Korean, which are agglutinative languages; Chinese, Thai, and Arabic, which are analytic languages; and English, which can be analytic or fusional language. These languages have different writing systems, including logogram (Chinese), syllabic system (Japanese), featural system (Korean), abugida (Thai), abjad (Arabic), and true alphabets (English and German).

The English corpora consist of The New York Times (Sandhaus, 2008), the Yelp Dataset<sup>1</sup>, and United States State of the Union addresses (1790 to 2018) divided

<sup>1</sup>[www.yelp.com/dataset](http://www.yelp.com/dataset)

into paragraphs<sup>2</sup>. The German data is from Ten Thousand German News Articles Dataset<sup>3</sup>. The Chinese corpora consist of the news articles from Chinanews<sup>4</sup>, restaurant reviews from Dianping<sup>5</sup>, and the movie reviews from Douban<sup>6</sup>. The Japanese data is drawn from the Webhose’s Free Datasets<sup>7</sup>. The Korean data is from the KAIST Corpus<sup>8</sup>. The Thai corpora consist of the news articles in Prachathai<sup>9</sup>, the restaurant reviews from Wongnai<sup>10</sup>, the BEST corpus<sup>11</sup>, and the Thai National Corpus (Aroonmanakun, 2007). The Arabic data is from the Antcorpus (Chouigui et al., 2017). Each corpus is divided into 75% training documents and 25% test documents (Table 3.1).

We use the text from the reduce version of Wikipedia dump for each language except English to train the  $\chi^2$ ,  $t$ , and frequency-based tokenizers. For English we use the Wiki103 dataset (Merity et al., 2016). English, German, Chinese, Japanese, Korean, Thai and Arabic documents are tokenized with NLTK (Bird, 2006), SoMaJo (Proisl and Uhrig, 2016), Stanford Word Segmenter (Tseng et al., 2005), Fugashi (McCann, 2020), KoNLPy (Park and Cho, 2014), Attacut (Chormai et al., 2020) and Camel-tools (Obeid et al., 2020) respectively. We construct a list of 50,000 top bigrams with the highest scores in each criterion. We then merge words in the input of the LDA with these lists.

We use the `gensim` (Řehůřek and Sojka, 2010) with the Continuous Bag-of-Word (CBOW) algorithm (Mikolov et al., 2013) to obtain word embeddings. For English, we lowercase and lemmatize data. To lemmatize a word is to remove the inflectional ending of the word, resulting in a lemma, or base form of the word. An example of lemmatizing is converting “studies” to just “study.” For Korean,

<sup>2</sup>[www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017](http://www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017)

<sup>3</sup>[github.com/tblock/10kGNAD](https://github.com/tblock/10kGNAD)

<sup>4</sup>[www.chinanews.com](http://www.chinanews.com)

<sup>5</sup>[github.com/zhangxiangxiao/glyph](https://github.com/zhangxiangxiao/glyph)

<sup>6</sup>[www.kaggle.com/utmhikari/doubanmovieshortcomments](http://www.kaggle.com/utmhikari/doubanmovieshortcomments)

<sup>7</sup>[webhose.io/free-datasets/japanese-news-articles/](http://webhose.io/free-datasets/japanese-news-articles/)

<sup>8</sup>[semanticweb.kaist.ac.kr/home/index.php/KAIST\\_Corpus](http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus)

<sup>9</sup>[github.com/PyThaiNLP/prachathai-67k](https://github.com/PyThaiNLP/prachathai-67k)

<sup>10</sup>[www.kaggle.com/c/wongnai-challenge-review-rating-prediction](http://www.kaggle.com/c/wongnai-challenge-review-rating-prediction)

<sup>11</sup>[thailang.nectec.or.th/downloadcenter](http://thailang.nectec.or.th/downloadcenter)

	Domains	Docs (K)	Tokens (M)	%Merged		
				CHI	T	FREQ
EN-NYTimes	News	53	0.7	1.64	12.71	12.72
EN-SOTU	Speeches	42	0.8	0.86	9.76	10.33
EN-Yelp	Restaurants	67	2.1	0.16	7.85	8.97
DE-10kGNAD	News	222	1.9	0.09	7.46	7.68
CN-Chinanews	News	49	0.8	0.00	11.61	11.64
CN-Dianping	Restaurants	40	0.8	0.01	2.82	2.80
CN-Douban	Movies	98	0.6	0.03	4.17	4.23
JA-JapanNews	News	528	3.6	21.74	21.95	21.85
KO-KAIST	Misc	20	0.2	19.82	20.71	21.27
TH-Prachathai	News	32	4.4	0.07	15.97	14.06
TH-Wongnai	Restaurants	40	1.2	0.00	8.52	6.09
TH-BEST	Misc	7	2.1	0.03	14.94	13.09
TH-TNC	Misc	4	1.0	0.03	13.65	12.00
AR-ANT	News	60	1.1	0.16	26.13	27.45

Table 3.1: The details of corpora we use in this study (Cheevaprawatdomrong et al., 2022)

Japanese, and Arabic, we lemmatize the data. For German, Chinese, and Thai, we do not do any normalization.

We use `MALLET` (McCallum, 2002) with the default hyperparameters to train and evaluate topic models with 10, 50, 100 topics. We run the experiment 3 times for each combination of corpus, type of retokenization (no retokenization,  $\chi^2$ ,  $t$  or frequency), and number of topics to compute the means of the normalized held-out likelihood and CBES, which is explained in chapter 3.2.

## 3.4 Results and Discussion

### Results

We first describe the overall picture of the results to show the similarity among all corpora in all languages and then discuss individual languages later about the unique behaviors they have which are different from the majority.

In general, the retokenization based on  $t$  and frequency significantly improve the normalized log-likelihood per token for English, German, Chinese, Japanese, Korean, and Arabic for all text collections and the number of topics except EN-Yelp, TH-BEST, and TH-TNC (Table 3.3). Among these two measures, frequency-based retokenization performs slightly better than  $t$  retokenization.

The  $\chi^2$  retokenization does not perform well in most languages, except for Japanese and Korean. This result is counterintuitive since  $\chi^2$  is a widely used measure in finding collocation. It shows that  $\chi^2$  based collocation might not be suitable for merging input of the LDA model.

The retokenization based on  $t$  and frequency also improves the coherence of the topic-keys (Table 3.3). After applying the retokenization to the input, topic-keys become more semantically coherent, and topics become more distinct. We see the improvement across all number of topics in English, Japanese, Korean, and Arabic corpora.

Similar to the results in normalized log-likelihood, we see the improvement across all types of collocation measures for Japanese and Korean. There could be some quality in morphology or typology of the two languages that make them benefit from the retokenization.

## English

English is the language that yields the results that agree with the overall results.  $\chi^2$ -based retokenization doesn't significantly improve both the normalized log-likelihood per token and the coherence of the topic keys, if not make them worse. One of the results worth mentioning is that the  $t$  and frequency based retokenization improve both the normalized log-likelihood per token and the coherence of the topic keys for all number of topics over all corpora except only for the 100 topics of Yelp. This exception could be because the domain of the Yelp, which is restaurant reviews, is quite different from that of the Wikipedia used to train the retokenizers.

## Chinese and Thai

What happens to Yelp in English can be observed more clearly in Chinese and Thai. The corpora in both languages contain text from many domains. Prachathai and Chinanews consist of news, Wongnai and Dianping contain restaurant reviews, Douban involves movie reviews, and in BEST and TNC there are various types of text, including novels.

While, in general,  $t$  and frequency based retokenization improve the normalized log-likelihood and the coherence of the topic-keys, they fail to increase the topic coherence in Dianping and Douban in Chinese over almost all number of topics. In Thai, They worsen the normalized log-likelihood of the 10 topics of BEST and TNC. The frequency based retokenizer also performs poorly over all topics of Wongnai. This is worth mentioning because in the languages with many corpora from various domains, we can clearly observe that the  $t$  and frequency based retokenizers do not perform as expected in corpora having content diverge from that of the Wikipedia used to train the retokenizers.

## Japanese and Korean

The results stand out in Japanese and Korean. While  $\chi^2$ -based retokenization doesn't significantly increase, if not decrease, both the normalized log-likelihood

per token in most languages, it performs well across the board in Japanese and Korean. Some characteristics in these languages could be the reason for this exception.

### German and Arabic

There are some results in German and Arabic worth mentioning. In Arabic, while  $\chi^2$ retokenization worsens the log-likelihood per token over all number of topics, it improves the topic coherence across the board. It is almost the opposite in German, where the same retokenization improves the log-likelihood per token over all number of topics but decreases the topic coherence in 10 and 50 topics of the German corpus. We do not see any clear explanation for this behavior. The  $\chi^2$ retokenization doesn't help in general, so it can introduce some noise to the data, but some characteristics in these languages may help support its performance in certain metrics.

## Discussion

### Chi-square Strategy

We observe different improvements from collocation measures, and  $\chi^2$  does not perform well in most languages. This could be due to the percentage of merged tokens during the retokenization (Table 3.1). The percentages of merged tokens using  $\chi^2$  measure are about one percent or below for many corpora, including English, Chinese, German, Arabic, and Thai corpora. This could generate the noise into the data that makes the results worse than the baseline in some cases.

On the other hand, we can see significantly higher percentages of merged tokens using  $t$  and frequency-based retokenization. They merge similar tokens across languages as demonstrated in Table 3.2. We see about 8%-15% of merging in English, German and Chinese, The highest percentages of merging, which is about 26%-27% are in Arabic. Japanese and Korean see about 20% of merging in all three types of retokenization strategies. All of these higher merging percentages



The reason why  $\chi^2$ retokenization merges fewer tokens than other measures could be the truncation of the top bigrams list. We limit the number of top bigrams to 50,000 for all three measures. However, the number of the bigrams that pass the hypothesis testing is relatively large, so many of the bigrams from the  $\chi^2$ measure are left unused. For example there are 3.73 million  $\chi^2$ collocations, much more than the 231 thousand  $t$  collocations in Thai for the same significance level  $\alpha = 0.005$ . The list of bigrams that pass the  $\chi^2$ testing also includes all of the top bigrams obtained from the  $t$  measure. Therefore, we could get at least the same percentage merged if we use all the bigrams that pass the  $\chi^2$ hypothesis testing. In addition, we also observe that the top 50,000 bigrams from  $\chi^2$ measures mostly contain rare words. These words don't appear frequently, but they make the list because they co-occur much more than they should do randomly. So when the percentage merge is very low when we use these rarely occurred top 50,000 bigrams from  $\chi^2$ measure.

The writing system or the morphology of the language can also account for the different merged percentages. For example, in English, we see specific named entities in the top 20  $\chi^2$ collocations and see compound nouns and common phrases in the  $t$  and frequency-based retokenizers (Figure 3.2).

## Influence of Merge Percentages

The models with higher merge percentages produce better normalized log-likelihood and CBES scores. That means they are better regarding the goodness of fit and topic-keys coherence. When considering merging strategies,  $t$  and frequency measures give better merge percentages, and when focusing on each language, the news corpora see higher merge percentages over other types of data, such as restaurant and movie reviews (Table 3.1). This could be because the news corpora are in a similar domain to that of the Wikipedia, which we use to build the top bigrams list.

There is a positive correlation between merge percentage and the improvement of the PTLT over the baseline word model. The correlation coefficient is 0.41, 0.77,

$\chi^2$ : dvenadsat apostolov, jormp jomp, malwae tweep, aboul gheit, achduth vesholom, adavari matalaku, adeste fideles, afforementionede oughtt, agoraf drws, aht urhgan, akanu ibiam, aksak maboul, alberthiene endah, alfava metraxis, alfonsas eidintas, allasani peddana, alteram partem, amantes clandestinos, amarin winitchai, amel oluna
$t$ : united states, new york, world war, km h, take place, miles km, los angeles, united kingdom, first time, high school, tropical storm, new zealand, war ii, video game, mph km, h mph, north america, air force, two years, peak number
frequency: united states, new york, world war, km h, take place, miles km, first time, los angeles, united kingdom, high school, tropical storm, new zealand, video game, war ii, mph km, two years, h mph, north america, air force, peak number
$\chi^2$ : うそ寒い 肌寒, ぎっこん ばったん, ざらり ぐらり, へへへへ へへへ, アウレオルス ボンバストゥス, アジ タケサカンバリン, アッシュアルク アルアウサト, アトミズム アドリアシン, アドリアシン アドリアマイシ ン, アルパイ オザラン, アワサカ ツマオ, イブリツモマブ チウキセタン, ウダヤン プラサッド, ウラマツ サ ミタロウ, エウグランディナ ロセア, エストラムスチン エストラサイト, オクタクロルテトラヒドロ メタノ フタラン, オドネ センデロル, オランバヤル ビャンバジャブ, クツミ ソクチュウ
$t$ : 年月, る 居る, 月日, る 事, 其の後, 成る 居る, 昭和年, 事出る, 年昭和, 於くり, 年年, 成る, 事有る, 事成 る, 使用る, 物有る, 存在る, 平成年, 第回, る 年
frequency: る 居る, 年月, 月日, る 事, る 年, 年年, 成る 居る, 居る 年, 其の後, 事有る, 昭和年, る, る 其の, 事 成る, 事出る, 年昭和, 有る 年, 成る, 使用る, 於くり
$\chi^2$ : 가넛 알원소, 가윗일 붓일, 가츠테루 우루샤, 가톨리콘 엠블, 갈끔 가실끔, 갈라람 알부담, 감민 월민, 감성 채 널@21, 갑복 갑규, 강첸 키송, 강취완 강취일, 강홍업 강효업, 강홍선 강홍익, 개영 케영, 개초향 거륵향, 개 튀의알 똥퍼먹는, 객렬액 겁렬액, 갤러 리@KCUA, 갤런에서 갤런으로, 거대유방증 대유방
$t$ : 적 인, 하다 수, 한 다, 위 한, 말 하다, 시작 하다, 사용 하다, 못 하다, 수 없다, 위치 한, 하다 않다, 사용 되다, 하 다 위해, 가지 고, 기도 하다, 일반 적, 되다 않다, 존재 하다, 기록 하다, 은 대한민국
frequency: 적 인, 하다 수, 하다 하다, 한 다, 사용 하다, 말 하다, 시작 하다, 하다 않다, 위 한, 못 하다, 수 없다, 위 치 한, 하다 위해, 하다는, 사용 되다, 기록 하다, 되다 않다, 하다 되다, 기도 하다, 활동 하다

Figure 3.2: The top 20 collocations from each strategy. (Cheevaprawatdomrong et al., 2022)

and 0.68 for all models with 10, 50, and 100 topics respectively. There is also a positive correlation between merge percentage and the improvement of CBES over the baseline word model. The correlation coefficient is 0.73, 0.76, and 0.79 for all models with 10, 50, and 100 topics respectively. This means the retokenizers that generalize well and recognize many collocations in the target corpora can better improve the result of the LDA models.

## Topic-keys

We found merged topic-keys in almost all topics when the models are merged with the  $t$  or raw frequency measures. We can also see that the meaning of the topic-keys from the merged models is more precise. For example, the meaning of the collocation “social security” is much more precise than the individual meaning of its components, which are the word “social” and “security.” In some cases, the meaning is totally different when breaking up the collocations. For example, the meaning of the collocation “คนเสื้อแดง” [kʰon s a d ɲ], which is a specific political

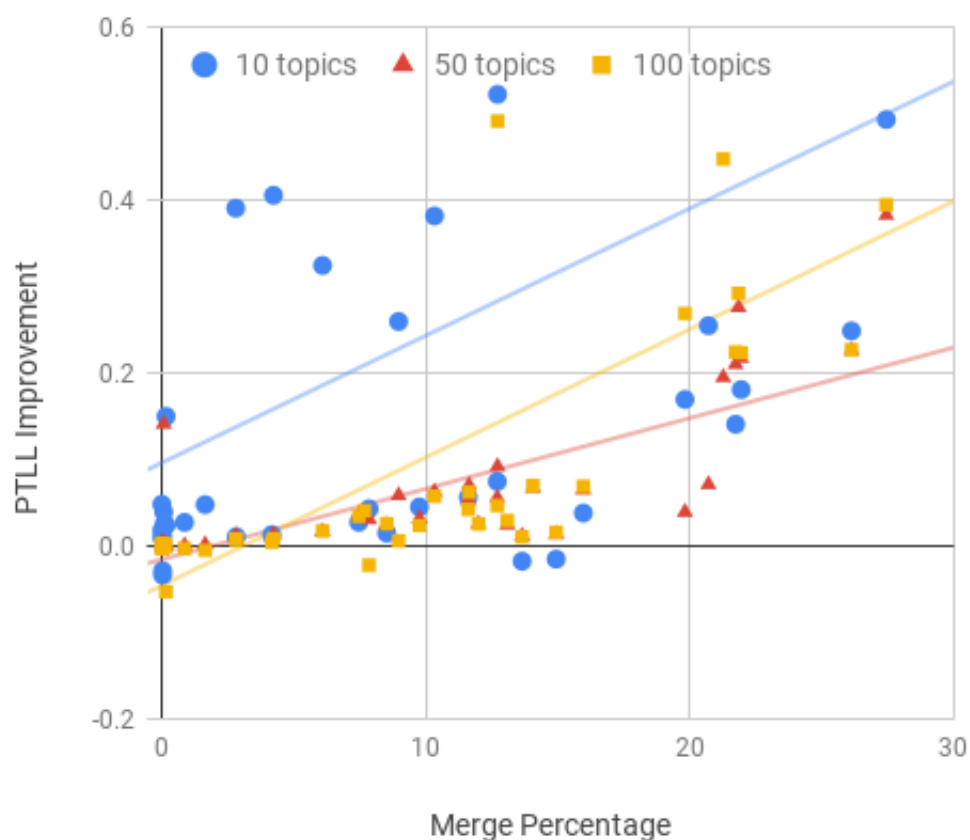


Figure 3.3: PTL improvement vs. merged percentage. (Cheevaprawatdomrong et al., 2022)

group in Thailand, can be totally lost when the collocation is broken into words, which are “คน” [k<sup>h</sup>on] (people) “เสื้อ” [s a] (shirt) “แดง” [d ɲ] (red).

However, some might feel that the topics from the merged and unmerged models look similar. That is because we look through the human perspective. We understand the language and understand the context of the topics. The computer and algorithms may not have this privilege, and they need explicit collocations to perform well in search and classification tasks.

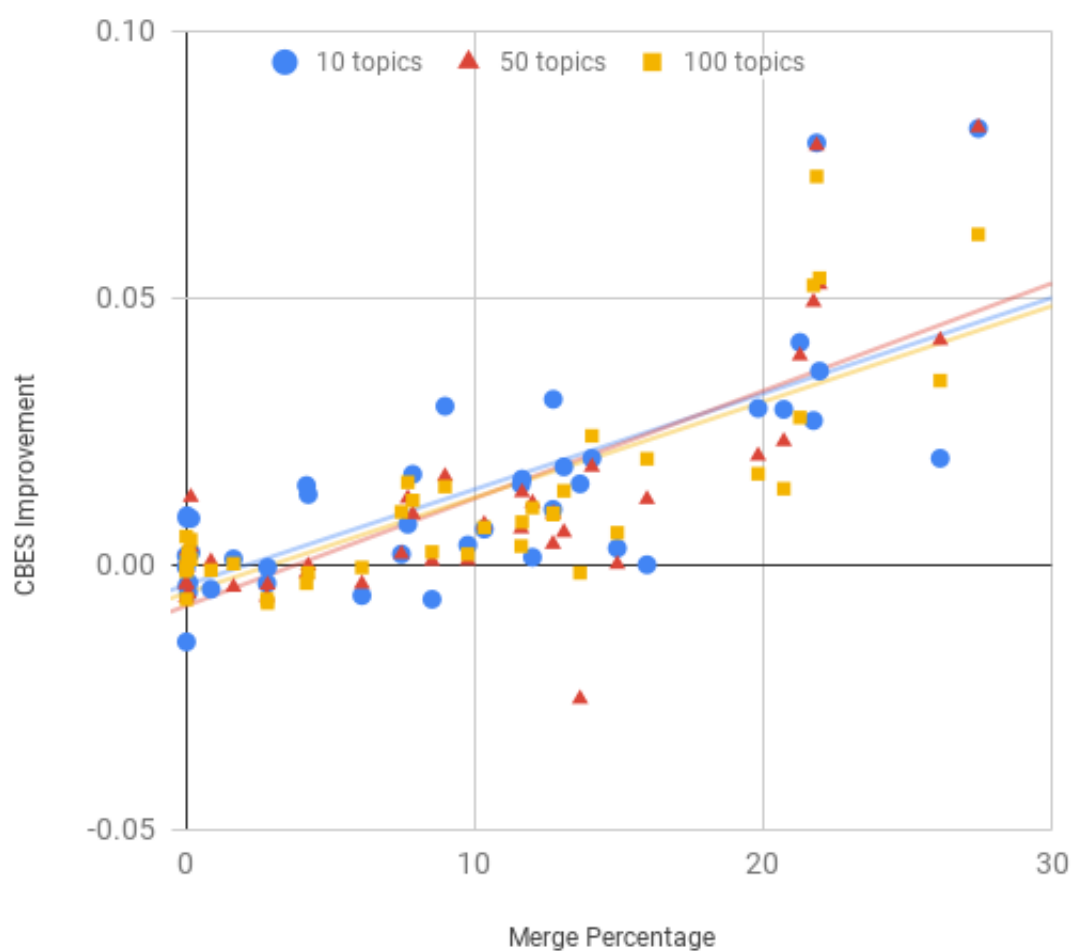


Figure 3.4: CBES improvement vs. merged percentage. (Cheevaprawatdomrong et al., 2022)

	$\chi^2-t$	$\chi^2$ -freq	$t$ -freq
English	8.90	7.78	74.87
German	0.00	0.00	83.06
Chinese	0.00	0.00	86.48
Japanese	29.06	22.60	73.34
Korean	10.56	7.34	71.95
Thai	0.22	0.06	67.25
Arabic	1.22	1.20	66.89

Table 3.2: The percentage of overlapping merged tokens between two methods of retokenization computed on the retokenization training data. (Cheevaprawatdomrong et al., 2022)

	10 topics				50 topics				100 topics			
	Word	$\chi^2$	$t$	freq	Word	$\chi^2$	$t$	freq	Word	$\chi^2$	$t$	freq
EN-NYTimes	.3646	.3675	.4119	<b>.4386</b>	.5214	.5225	.5766	<b>.6128</b>	.5588	.5533	.6050	<b>1.0492</b>
EN-SOTU	.2699	.2660	.2967	<b>.3145</b>	.3809	.3809	.4122	<b>.4430</b>	.4135	.4101	.4367	<b>.4705</b>
EN-Yelp	.1597	.1607	.1833	<b>.2021</b>	.2589	.2599	.2893	<b>.3169</b>	.3357	.2822	.3130	<b>.3412</b>
DE-10kGNAD	.4982	.5001	.5233	<b>.5251</b>	.7272	.7272	.7622	<b>.7651</b>	.7784	.7809	.8122	<b>.8188</b>
CN-Chinanews	.5033	.5046	.5510	<b>.5592</b>	.7647	.766	.8170	<b>.8344</b>	.8427	.8394	.8847	<b>.9044</b>
CN-Dianping	.2557	.2574	.2644	<b>.2659</b>	.3899	.3906	.3965	<b>.4013</b>	.4188	.4212	.4255	<b>.4263</b>
CN-Douban	.2966	.2955	.3076	<b>.3092</b>	.4048	.4073	.4144	<b>.4173</b>	.4294	.4301	.4332	<b>.4374</b>
JA-JapanNews	.4540	<b>.7803</b>	.5942	.6342	.7173	.9268	.9339	<b>.9926</b>	.8088	1.0325	1.0316	<b>1.1003</b>
KO-KAIST	.2901	<b>1.0315</b>	.4589	.5442	.6446	.6833	.7152	<b>.8390</b>	.4755	.7437	<b>1.3443</b>	.9221
TH-Prachathai	.4367	.4331	<b>.4756</b>	.4743	.7052	<b>.8458</b>	.7699	.7719	.7854	.7854	.8537	<b>.8548</b>
TH-Wongnai	.2048	.2013	<b>.2225</b>	.2192	.3237	.3222	<b>.3472</b>	.3399	.3467	.3463	<b>.3720</b>	.3636
TH-BEST	<b>.6995</b>	<b>.6995</b>	.6704	.6838	.9148	.9190	.9279	<b>.9389</b>	.9812	.9819	.9967	<b>1.0100</b>
TH-TNC	.7420	<b>.7422</b>	.7079	.7239	.9969	.9952	1.0079	<b>1.0219</b>	1.0508	1.0473	1.0608	<b>1.0758</b>
AR-ArabNews	.3183	.3152	.4676	<b>.5663</b>	.4923	.4913	.7175	<b>.8742</b>	.5417	.5409	.7681	<b>.9355</b>

	10 topics				50 topics				100 topics			
	Word	$\chi^2$	$t$	freq	Word	$\chi^2$	$t$	freq	Word	$\chi^2$	$t$	freq
EN-NYTimes	.0143	.0153	.0246	<b>.0453</b>	-.0582	-.0625	-.0544	<b>-.0487</b>	-.0876	-.0875	-.0783	<b>-.0780</b>
EN-SOTU	.0034	-.0013	.0070	<b>.0100</b>	-.0602	-.0597	-.0595	<b>-.0527</b>	-.0812	-.0823	-.0793	<b>-.0743</b>
EN-Yelp	-.0634	-.0548	-.0465	<b>-.0337</b>	-.1117	-.1085	-.1023	<b>-.0952</b>	-.1299	-.1290	-.1179	<b>-.1153</b>
DE-10kGNAD	-.0209	-.0244	-.0190	<b>-.0134</b>	-.0804	-.0860	-.0785	<b>-.0680</b>	-.0753	-.0730	-.0655	<b>-.0599</b>
CN-Chinanews	.0002	.0018	.0152	<b>.0162</b>	-.0523	-.0559	-.0456	<b>-.0388</b>	-.0699	-.0712	-.0665	<b>-.0620</b>
CN-Dianping	<b>-.0708</b>	-.0854	-.0714	-.0744	<b>-.1278</b>	-.1316	-.1317	-.1339	<b>-.1373</b>	-.1439	-.1446	-.1439
CN-Douban	-.0226	-.0140	<b>-.0078</b>	-.0095	<b>-.0847</b>	-.0854	-.0864	-.0850	<b>-.1037</b>	-.1041	-.1073	-.1053
JA-JapanNews	-.0925	-.0655	-.0562	<b>-.0133</b>	-.1503	-.1010	-.0977	<b>-.0716</b>	-.1644	-.1120	-.1106	<b>-.0915</b>
KO-KAIST	-.0608	-.0315	-.0317	<b>-.0191</b>	-.0895	-.0691	-.0664	<b>-.0503</b>	-.0868	-.0698	-.0726	<b>-.0592</b>
TH-Prachathai	<b>-.0039</b>	-.0092	-.0040	.0160	-.0806	-.0797	-.0684	<b>-.0623</b>	-.1137	-.1121	-.0939	<b>-.0896</b>
TH-Wongnai	<b>-.0667</b>	-.0672	-.0733	-.0726	-.1468	-.1530	-.1462	-.1505	-.1761	-.1709	-.1738	-.1767
TH-BEST	-.0278	-.0187	-.0248	<b>-.0095</b>	-.0987	-.0977	-.0987	<b>-.0927</b>	-.1145	-.1153	-.1086	<b>-.1007</b>
TH-TNC	-.0284	-.0324	<b>-.0133</b>	-.0271	-.1079	-.1053	-.1332	<b>-.0964</b>	-.1281	-.1274	-.1297	<b>-.1175</b>
AR-ArabNews	-.0695	-.0673	-.0496	<b>.0124</b>	-.1255	-.1129	-.0834	<b>-.0434</b>	-.1355	-.1309	-.1010	<b>-.0735</b>

Table 3.3: Normalized unigram log-likelihood per token (top) and Concatenation-based Embedding Silhouette (CBES) scores (bottom) for between the baseline and retokenization models:  $\chi^2$ , textit, and raw frequency. (Cheevaprawatdomrong et al., 2022)

## Chapter IV

### CONCLUSION

In this research, we show that merging input tokens of the LDA can improve the statistical fit of the model, and results in the topics that are more coherent and more distinct. We also found that the percentage of merging has a positive impact on both goodness of fit and coherence results. The study shows that  $t$  statistics and raw frequency strategies are better than the  $\chi^2$  measure when we want to merge LDA input because the  $\chi^2$  measure focuses on rare named entities which do not merge well in general documents. By retokenizing with  $t$  statistics and frequency measure, we get the input with noun phrases in the topic keys that could help better understanding of the topics and make these topics more semantically precise for the downstream tasks. We also found different merging behavior, and thus varying results, among types of languages. This may be due to their unique morphological typology and writing system.

## REFERENCES

- Aroonmanakun, W. 2007. Creating the thai national corpus. MANUSYA: Journal of Humanities 10.3 (2007): 4–17.
- Bin, G., HE, C.-h., HU, S.-z., and Cheng, G. 2018. Chinese news hot subtopic discovery and recommendation method based on key phrase and the lda model. DESTech Transactions on Engineering and Technology Research .ecar (2018):
- Bird, S. 2006. Nltk: the natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pp. 69–72. :
- Blei, D. M. 2012. Probabilistic topic models. Communications of the ACM 55.4 (2012): 77–84.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. Journal of machine Learning research 3.Jan (2003): 993–1022.
- Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL (2009): 31–40.
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. (ed.), Advances in Neural Information Processing Systems, volume 22. : Curran Associates, Inc.
- Cheevaprawatdomrong, J., Schofield, A., and Rutherford, A. T. 2022. More than words: Collocation tokenization for latent dirichlet allocation models. In Findings of the Association for Computational Linguistics. :
- Chormai, P., Prasertsom, P., Cheevaprawatdomrong, J., and Rutherford, A. 2020. Syllable-based neural thai word segmentation. In Proceedings of the 28th International Conference on Computational Linguistics, pp. 4619–4637. :

- Chouigui, A., Khiroun, O. B., and Elayeb, B. 2017. Ant corpus: an arabic news text collection for textual classification. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 135–142. :
- Dawson, H., Phelan, M., et al. 2016. Language files: Materials for an introduction to language and linguistics. The Ohio State University Press.
- Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19.1 (1993): 61–74.
- El-Kishky, A., Song, Y., Wang, C., Voss, C., and Han, J. 2014. Scalable topical phrase mining from text corpora. arXiv preprint arXiv:1406.6312 (2014):
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. 1995. Markov chain Monte Carlo in practice. CRC press.
- Griffiths, T. and Steyvers, M. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101 (2004): 5228–5235.
- Lau, J. H., Baldwin, T., and Newman, D. 2013. On collocations and topic models. ACM Transactions on Speech and Language Processing (TSLP) 10.3 (2013): 1–14.
- Li, B., Yang, X., Zhou, R., Wang, B., Liu, C., and Zhang, Y. 2018. An efficient method for high quality and cohesive topical phrase mining. IEEE Transactions on Knowledge and Data Engineering 31.1 (2018): 120–137.
- Lindsey, R., Headden, W., and Stipicevic, M. 2012. A phrase-discovering topic model using hierarchical pitman-yor processes. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 214–222. :
- Manning, C. and Schütze, H. 1999. Foundations of statistical natural language processing. MIT press.



- May, C., Cotterell, R., and Van Durme, B. 2016. An analysis of lemmatization on topic models of morphologically rich language. arXiv preprint arXiv:1608.03995 (2016):
- McCallum, A. K. Mallet: A machine learning for language toolkit, 2002.
- McCann, P. 2020. fugashi, a tool for tokenizing Japanese in python. In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pp. 44–51. Online: Association for Computational Linguistics.
- Merity, S., Xiong, C., Bradbury, J., and Socher, R. 2016. Pointer sentinel mixture models.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013):
- Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. 2011. Optimizing semantic coherence in topic models. In Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 262–272. :
- Mood, A. M., Graybill, F. A., and Boes, D. C. 1974. Special parametric families of univariate distributions. Introduction to the theory of statistics, 3rd ed. McGraw-Hill, New York (1974): 119–120.
- Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. 2020. CAMEL tools: An open source python toolkit for Arabic natural language processing. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 7022–7032. Marseille, France: European Language Resources Association.
- Park, E. L. and Cho, S. 2014. Konlpy: Korean natural language processing in python. In Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. Chuncheon, Korea:

- Pearson, K. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50.302 (July 1900): 157–175.
- Proisl, T. and Uhrig, P. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pp. 57–62. Berlin: Association for Computational Linguistics (ACL).
- Řehůřek, R. and Sojka, P. 2010. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. Valletta, Malta: ELRA.
- Rogers, H. 2005. Writing systems: A linguistic approach, volume 18. Blackwell publishing.
- Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20 (1987): 53–65.
- Sandhaus, E. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia 6.12 (2008): e26752.
- Schofield, A. and Mimno, D. 2016. Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics 4 (2016): 287–300.
- Schofield, A., Magnusson, M., and Mimno, D. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 432–436. :
- Student. 1908. The probable error of a mean. Biometrika (1908): 1–25.

- Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., and Manning, C. D. 2005. A conditional random field word segmenter for sighthan bakeoff 2005. In Proceedings of the fourth SIGHAN workshop on Chinese language Processing. :
- Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. 2009. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, p. 1105–1112. New York, NY, USA: Association for Computing Machinery.
- Wang, M., Zhao, B., and Huang, Y. 2016. Ptr: phrase-based topical ranking for automatic keyphrase extraction in scientific publications. In International Conference on Neural Information Processing, pp. 120–128. :
- Yu, Z., Johnson, T. R., and Kavuluru, R. 2013. Phrase based topic modeling for semantic information processing in biomedicine. In 2013 12th International Conference on Machine Learning and Applications, volume 1, pp. 440–445. :

# Appendix I

## TOP BIGRAMS

frequency	united states, new york, world war, km h, take place, miles km, first time, los angeles, united kingdom, high school, tropical storm, new zealand, video game, war ii, mph km, two years, h mph, north america, air force, peak number, number one, years later, music video, year old, km mi, york city, inch mm, follow year, days later, next day, positive review, become first, prime minister, final fantasy, civil war, also know, metres ft, new jersey, york times, long tons, take part, tropical depression, three years, studio album, two days, san francisco, tropical cyclone, first two, television series, science fiction
t-statistics	united states, new york, world war, km h, take place, miles km, los angeles, united kingdom, first time, high school, tropical storm, new zealand, war ii, video game, mph km, h mph, north america, air force, two years, peak number, music video, km mi, year old, years later, inch mm, york city, number one, days later, follow year, positive review, next day, prime minister, final fantasy, civil war, metres ft, new jersey, york times, long tons, also know, tropical depression, become first, studio album, san francisco, tropical cyclone, three years, two days, take part, science fiction, television series, hip hop
Chi-square	dvenadsat apostolov, jormp jomp, malwae tweep, aboul gheit, achduth vesholom, adavari matalaku, adeste fideles, afforementionede oughtt, agoraf drws, aht urhgan, akanu ibiam, aksak maboul, alberthiene endah, alfava metraxis, alfonsas eidintas, allasani peddana, alteram partem, amantes clandestinos, amarin winitchai, amel oluna, anaganaga dheerudu, analysin infinitorum, anemer bangkong, angraecum sesquipedale, anjaana anjaani, annamaria pinazzi, anquan boldin, antikes seewesens, anthrushes antpittas, anubhavam pudhumai, archinome jasoni, arly singou, arsens miskarovs, aschwin wildeboer, athisaya piravi, athous lavrensis, athrotaxis selaginoides, atikaya akshayakumara, audemars piguet, austinograea rodriguezensis, avata nirodhana, ayaan hirsi, ayoub odisho, bachao andolan, baeospora myosura, baldrs draumar, barang tiada, bathys ryax, beauveria bassiana, belajske poljice

Figure A.1: The top 50 collocations from English Wikipedia. (Cheevap-rawatdomrong et al., 2022)

frequency	Im Jahr, New York, In Jahren, Zweiten Weltkrieg, Vereinigten Staaten, Im Jahre, Familienname Personen, Jahre später, University of, of the, Alter Jahren, selben Jahr, Es gibt, Darüber hinaus, Der Ort, Die erste, Die Gemeinde, Am Mai, Am Januar, de la, Am Juli, zwei Jahre, Am März, Am April, In Saison, Am Juni, Am Oktober, Der Film, Einwohnern Stand, Olympischen Spielen, Name Personen, Am September, Der Name, In Zeit, Ende Jahrhunderts, Ersten Weltkrieg, Jahr später, Am Dezember, Am November, ersten Mal, Die Stadt, Am August, Los Angeles, Mitte Jahrhunderts, Im Jahrhundert, Im selben, Die Liste, Des Weiteren, Frankfurt Main, Jahre lang
t-statistics	Im Jahr, New York, In Jahren, Zweiten Weltkrieg, Vereinigten Staaten, Im Jahre, Familienname Personen, University of, Jahre später, of the, Alter Jahren, selben Jahr, Darüber hinaus, Es gibt, Der Ort, Am Mai, de la, Am Januar, Am Juli, Am März, Einwohnern Stand, Am April, Olympischen Spielen, Name Personen, Am Juni, Am Oktober, zwei Jahre, Ende Jahrhunderts, In Saison, Ersten Weltkrieg, Am September, Der Film, Die erste, Der Name, Am Dezember, ersten Mal, Jahr später, Die Gemeinde, In Zeit, Am November, Los Angeles, Am August, Mitte Jahrhunderts, Im selben, Des Weiteren, Frankfurt Main, Jahre lang, Im Jahrhundert, Hälfte Jahrhunderts, Am Februar
Chi-square	Budleigh Salterton, dagestanisch punjabisch, punjabisch sikhistisch, thesen temperamente, AAORRAC OARC, AARNLPNICK IPAVGRCQFT, AARNLPNICN IPNIGACPFR, ABACABA Kettenrondo, ABBATIAE LUDGERI, ABBYY FineReader, ABDAC Alkylteil, ABFF Studiopremiere, ABGESCHIDEN IHRES, ABGEW ENDET, ABOREA Heredium, ABSCONDITE ELEMOSINAM, ABSDORF ALBERNDORF, ACEP Lehrgangskonzept, ACOL Adeguamento, ACOMA Volksfrauen, ACRS Proliferationsfragen, ACTOR UNDERGOER, ADCs DACs, ADDUI SUBUI, ADFECTUS VULNERE, ADMIRABILE OPVS, ADOLF FRIEDRICHS, ADOLPH KOHUT, ADRW Naturpower, ADST Kantenblöcke, AELIVS SEVERINVS, AENEAS ORESTES, AENK Dimoulas, AEPB LRVF, AErosol RObotic, AFDRU Minensuchhunde, AFFINOMICS CAGEKID, AFIL Systemherstellers, AFYHW KXAXF, AGADF AAADG, AGITATED SCREAMS, AGOK AGOFF, AHENEUS BONA, AICTO Elgazala, AIIMS Gursaran, AIRAC AMDT, AIRNORTHWEST NAVNORTHWEST, AIUS Agrosorsurs, AJK Programmtagung, AJMER Ajaymeru

Figure A.2: The top 50 collocations from German Wikipedia. (Cheevap-rawatdomrong et al., 2022)

frequency	面积 平方公里, 人民 共和国, 人口 密度, 中华 人民, 人口 人口, 密度 平方公里, 总 面积, 全 国, 海拔 高度, 高度 米, 中华 民国, 中国 大陆, 行政 区划, 市镇 总, 负责 管辖, 政治 人物, 总 人口, 有 限 公 司, 平 方 公 里 人 口, 第 一 次, 位 于 国, 以 下 地 区, 号 线, 行 政 单 位, 第 二 次, 乡 镇 级 行 政, 下 辖 乡 镇 级, 平 方 公 里 海 拔, 米 人 口, 人 口 普 查, 首 次, 中 国 人 民, 变 化 图 示, 面 积 平 方 千 米, 下 辖 以 下, 人 口 变 化, 平 方 公 里 总, 中 国 共 产 党, 文 物 保 护, 保 护 单 位, 位 于 美 国, 次 世 界, 世 界 大 战, 两 种, 人 民 政 府, 人 民 解 放 军, 美 国 人 口, 属 下 种, 男 性 女 性, 位 国
t-statistics	面积 平方公里, 人民 共和国, 人口 密度, 中华 人民, 密度 平方公里, 人口 人口, 总 面积, 全 国, 海拔 高度, 高度 米, 中华 民国, 中国 大陆, 行政 区划, 市镇 总, 负责 管辖, 政治 人物, 有 限 公 司, 总 人 口, 平 方 公 里 人 口, 第 一 次, 位 于 国, 以 下 地 区, 号 线, 行 政 单 位, 乡 镇 级 行 政, 下 辖 乡 镇 级, 第 二 次, 平 方 公 里 海 拔, 人 口 普 查, 变 化 图 示, 米 人 口, 面 积 平 方 千 米, 下 辖 以 下, 首 次, 人 口 变 化, 中 国 人 民, 平 方 公 里 总, 中 国 共 产 党, 文 物 保 护, 保 护 单 位, 位 于 美 国, 世 界 大 战, 次 世 界, 人 民 解 放 军, 人 民 政 府, 两 种, 属 下 种, 男 性 女 性, 中 共 中 央, 辐 鳍 鱼
Chi-square	内 胚 窦 瘤, 一 举 千 里 羽 翻 已 就, 一 举 而 三 善 备 傅 嵩 林, 一 举 而 空 朔 庭 至 乃 追 奔 稽, 一 二 三 弦 宏 松, 一 二 如 二 一 一 如 一, 一 二 言 官 上 疏 极 谏, 一 二 魁 儒 负 劬, 一 介 匹 夫 哀 家, 一 价 金 三 价 金, 一 任 挝 挝 多 挝, 一 传 十 十 传 百, 一 作 慈 溢 戴, 一 例 仿 银 裹 金, 一 八 五 零 年 一 九 零 零 年, 一 分 五 厘 九 毫 二 秒 七 忽 胸 数, 一 切 冗 费 尽 裁, 一 券 无 伏, 一 副 烂 牌 嘟 著 嘴, 一 匹 缣 余 悉 分 士 伍, 一 号 球 二 号 球, 一 名 丞 或 丞, 一 名 仲 字 敬 真, 一 名 会 字 子 禽, 一 名 俊 字 述 道, 一 名 卷 作 盆, 一 名 芑 字 际 唐, 一 名 菱 黄 帝 云, 一 名 胡 张 騫 使, 一 名 隐 字 鸿 隐, 一 吐 欲 罢 不 能 者, 一 吹 八 伊 嘎 其 声, 一 品 云 九 品 云 九, 一 品 区 一 品 乡, 一 大 一 小 虾 螯, 一 女 嫁 刘 鄴 子, 一 女 施 涌 乐, 一 如 隔 昨 才, 一 婷 之 妹 赵 又 玲 杜 薇 饰, 一 字 其 明 号 德 云, 一 字 尔 器 初 字 征 鸣, 一 字 揭 夫 号 药 身, 一 宫 町 津 名 町, 一 宫 町 各 町 御 坂 町 各 町, 一 尺 六 寸 胃 纤 曲 屈, 一 幅 醉 杨 妃 图, 一 度 遣 他 戎, 一 座 皆 惊 金 叵 罗, 一 得 号 鸣 冈, 一 念 证 省 识

Figure A.3: The top 50 collocations from Chinese Wikipedia. (Cheevap-rawatdomrong et al., 2022)

frequency	<p>為る 居る, 年月, 月日, 為る 事, 為る 年, 年年, 成る 居る, 居る 年, 其の 後, 事 有る, 昭和 年, 為る 為, 為る 其の, 事 成る, 事 出来る, 年 昭和, 有る 年, 様 成る, 使用 為る, 於くり, 物 有る, 為る 此の, 為る 物, 存在 為る, 平成 年, 居る 事, 化 為る, 第 回, 成る 年, 年 平成, 同年 月, 年代, 発表 為る, 行う 居る, 有る 事, 居る 此の, 此れ 等, 日 年, 為る 言う, 為る 様, 放送 為る, 居る 其の, 有る 此の, 登場 為る, 合衆 国, 発売 為る, 明治 年, 為る 月, 開催 為る, 為る 此れ</p>
t-statistics	<p>年月, 為る 居る, 月日, 為る 事, 其の 後, 成る 居る, 昭和 年, 事 出来る, 年 昭和, 於くり, 年年, 様 成る, 事 有る, 事 成る, 使用 為る, 物 有る, 存在 為る, 平成 年, 第 回, 為る 年, 年 平成, 同年 月, 為る 為, 居る 年, 年代, 此れ 等, 発表 為る, 化 為る, 合衆 国, 行う 居る, アメリカ 合衆, 明治 年, 選手 権, 為る 其の, 高等 学校, 第 二, 登場 為る, 第 一, 為る 物, 開催 為る, 委員会, 居る 此の, 株式 会社, 設置 為る, 発売 為る, 居る 事, 其の 他, 小 学校, 為る 此の, 日 本</p>
Chi-square	<p>うそ寒い 肌寒, ぎっこん ばったん, ざらりぐらり, へへへへ へへへ, アウレオールス ボンバストゥス, アジタ ケーサカン バリン, アッシュアルク アルアウサト, アトミズム アドリアシン, アドリアシン アドリアマイシン, アルパイ オザラン, アワサカ ツマオ, イブリツモマブ チウキセタン, ウダヤン プラサッド, ウラ マツ サミタロウ, エウグランディナ ロセア, エストラムスチン エストラサイト, オクタクロルテトラヒドロ メタノフタラン, オドネ センデロール, オランバヤル ビャンバジヤブ, クツミ ソクチュウ, クロタン シャビニョル, クロルプロパミド グリベンクラミド, シャルワール カミーズ, ストラクチュアル オーナメンタル, ダイゴイチノキリチョウ ニノキリチョウ, チョウタツ ハンキョウ, ツアッキ アンニーナ, デンドロ セネキオ, トミロン フロモックス, トヨシタ ナラヒコ, トリクロピル スルホメツロン, ドウイドウ モウ, ナカイシキリチョウ ニシイシキリチョウ, ニョゲン ジンクウ, パクダ カッチャーヤナ, パルミジャーノ レッジャーノ, パーラミター フリダヤ, フルドロコルチゾン フロリネフ, プリン ペラン ガナトン, プレシオモナス シゲロイデス, ペンディメタ リンリニューロン, マルーシュカ デートメルス, ルツィオ クロ チェッティ, レザール フロリサン, 一網 打尽, 一罰 百戒, 偕 老 同穴, 克雪 利雪, 前唇 後唇, 北槎 聞略</p>

Figure A.4: The top 50 collocations from Japanese Wikipedia. (Cheevap-rawatdomrong et al., 2022)

frequency	적 인, 하다 수, 하다 하다, 한 다, 사용 하다, 말 하다, 시작 하다, 하다 않다, 위 한, 못 하다, 수 없다, 위치 한, 하다 위해, 하다는, 사용 되다, 기록 하다, 되다 않다, 하다 되다, 기도 하다, 활동 하다, 하다 이다, 존재 하다, 위 하다, 하다 은, 가지 고, 은 대한민국, 속 하다, 일반 적, 하다는, 참여 하다, 하다 이후, 이용 하다, 하다 위, 늘다 미국, 주장 하다, 사망 하다, 는 명, 명 이다, 관 한, 하다 되어다, 발생 하다, 발표 하다, 위치 하다, 등장 하다, 인구 는, 뜻 은, 차지 하다, 출연 하다, 배우 이다, 때문 이다
t-statistics	적 인, 하다 수, 한 다, 위 한, 말 하다, 시작 하다, 사용 하다, 못 하다, 수 없다, 위치 한, 하다 않다, 사용 되다, 하다 위해, 가지 고, 기도 하다, 일반 적, 되다 않다, 존재 하다, 기록 하다, 은 대한민국, 활동 하다, 늘다 미국, 참여 하다, 관 한, 속 하다, 이용 하다, 사망 하다, 주장 하다, 는 명, 명 이다, 인구 는, 발생 하다, 등장 하다, 배우 이다, 뜻 은, 발표 하다, 지정 되어다, 발견 되다, 차지 하다, 출연 하다, 때문 이다, 위 하다, 참가 하다, 이기도 하다, 볼 수, 출전 하다, 생각 하다, 의미 하다, 영화 이다, 축구 선수
Chi-square	가닛 알흰소, 가윗일 붓일, 가즈테루 우루샤, 가톨리콘 앰블, 갈꾸 가실꾸, 갈라람 알부담, 감민 월민, 감성채 널 @21, 갑복 갑규, 강첸 키송, 강취완 강취일, 강홍업 강효업, 강홍선 강홍익, 개영 케영, 개초항 거륵항, 개튀의알 똥퍼먹는, 객렬액 겁렬액, 갤러 리@KCUA, 갤런에서 갤런으로, 거대유방증 대유방, 거독 옥책안, 거른대 낭텡, 거릿기 사랏, 거칸 생건, 걸룬 룬미, 검류혼 목정균, 겹부 겹보, 게림 토샤빔, 겐된 최펠, 겐소미 콘스탄초, 격근 액진, 견고라스 아마루스, 견군공 응우옌폭홍까이, 견망 견둥, 견취사 계취사, 결갑목 골갑목, 겹실 꿩었으, 경계유대 소연유대, 경산현 압량군, 계금취신계 차실집취신계, 계반 밥뚜경, 계혁주 계형봉, 고가레 타오키야, 고련기신 고련제옥, 고련제옥 낙상잡주, 고류지인 집법지인, 고브 메빌렉 관갱, 고상점 액상점, 고킨 신고킨, 고행상 장행상

Figure A.5: The top 50 collocations from Korean Wikipedia. (Cheevap-rawatdomrong et al., 2022)

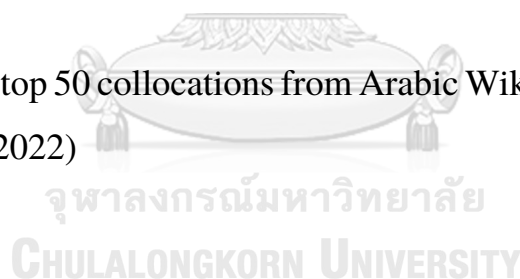


frequency	วันที่, ได้รับ, ในปี, ในการ, ที่มี, ทำให้, มีการ, ต่อมา, อยู่ใน, มีความ, เมื่อวัน, ที่จะ, หลังจาก, ซึ่งเป็น, จาก, การ, รับ, การ, เป็น, ผู้, มาจาก, จะมี, ไม่มี, เป็นการ, ในช่วง, ในวัน, เป็น, ที่, ที่เป็น, จาก, นั้น, การ, ศึกษา, และ, มี, นอกจาก, ไม่ได้, การ, แข่งขัน, เนื่องจาก, และ, การ, ว่า, เป็น, โดยมี, ที่ได้, และ, ได้, และเป็น, ครั้ง, แรก, อื่น ๆ, ขึ้น, ใน, เกิด, ขึ้น, จะ, เป็น, ครั้งที่, ได้, มี, เป็น, นัก, เดียว, กัน, ต่าง ๆ, ทำงาน, หนึ่งใน
t-statistics	ได้รับ, วันที่, ในปี, ทำให้, ต่อมา, มีการ, ที่มี, ในการ, เมื่อ, วัน, หลังจาก, อยู่ใน, มีความ, รับ, การ, ซึ่งเป็น, ในช่วง, มา, จาก, เป็น, ผู้, ที่จะ, นอกจาก, ไม่มี, การ, ศึกษา, เนื่องจาก, จาก, การ, จาก, นั้น, การ, แข่งขัน, จะมี, อื่น ๆ, ครั้ง, แรก, ในวัน, เกิด, ขึ้น, เดียว, กัน, ต่าง ๆ, ทำงาน, ไม่ได้, เรียกว่า, เกี่ยวกับ, ตะวัน, ออก, ผลงาน, ตั้ง, อยู่, โรงเรียน, ตั้ง, กล่าว, เป็น, นัก, ภายใน, มากกว่า, ครั้งที่, เข้าร่วม, ตะวัน, ตก, โดยมี, ออก, มา, ของ, เขา
Chi-square	คลอแรม เฟนิคอล, นิกายม หีศาจ, พีร์วัส แสงโพธิ์รัตน์, มหาสง ขมิเกะ, ยุราน อเมเทรีย, รติพงษ์ ภูมาลี, ศุภมาศ พะหุโล, อัคราภิธาน ศรีบัท, อาแซตแอส แซญสตอคควา, เอศวต มาเนท์, โค้ด ออฟบลู, กกกร เบญจาทิกุล, กกกอ หล้าคมบาง, กกปกหอม พิซซซซซซ, กกม่วงซี ถนนมะขามหวาน, กกลุ่มมอร์นิ่งมูเซะไ, กกะ อีสาร์, กัก บฟู, กฤดาลงกรณ ชัตติยจก, กกเฮือ ชี้ซัน, กคาลิมา มุตลาอินเท, กงฉายาปักข่าย ขอทานทาง, กงชู จูเล่ย์, กงตา เฉย์, กงปาออ เรียบ, กงม้า คอนม้า, กงสต็องซ์แห่ง เบียร์ริง, กงสุลลิเบโร แองเจลซซี, กงงานเซย, กงชาฐานิเยสุ ธเมมสุ, กงโนะริโซโตะ เก็งจิสึ, กงไซเซา อาไลชู, กงชณช เกียนชินวรา, กงชธรรมคน ธันย์วุฒิ, กงชมน คาร์ส, กงชายหนี, กงช้อชตอฟ ซกูปี้แซฟสกี, กงชลาयरัน, กงจาริคประเพณิ, กงทระกุลเดอ แฟคโต, กงสเตฟาน โบล์ทชมาน, กงกติ การมารตรฐาน, กงม หาเถรสมาคม, กงมล เฑียรบาลราชวงศ์, กงนาท และปกุช, กงตปญโญ วัดป่าผาเทพนิมิตร, กงต ดอย, กงตญชลี ชัยรัตน์ศิริพงศ์, กงตโม จปุกคโล, กงตันฉบับเดิ

Figure A.6: The top 50 collocations from Thai Wikipedia. (Cheevaprawat-domrong et al., 2022)

frequency	<p>ولاية مُتَّجِد، وُلِدَ فِي، فِي وِلَايَةٍ، نِسْبَةٌ مِنْ، بَلَغَ عَدَدٌ، عَدَدٌ سَاكِنٍ، عَدِيدٌ مِنْ، مِنْ قَبْلِ، لَاعِبِ كُرَّةٍ، ثَامِنٌ عَشْرَةَ، أَمْرِيكِيٌّ، وُلِدَ، كُرَّةٌ قَدَّمَ، هُوَ لَاعِبٌ، تَوَقَّى فِي، مِنْ خِلَالِ، نَسَمَةٌ حَسَبٌ، أَكْثَرُ مِنْ، وَقَعَ فِي، فِي ذَلِكَ، كُرَّةٌ قَدَّمَ، فِي مَدِينَةٍ، فِي مَمْلَكَةٍ، فِي مَنطِقَةٍ، رَعْمٌ مِنْ، حَسَبَ تَعْدَادِ، فِي فِي، مِنْ أَصْلٍ، إِضَافَةٌ إِلَى، لِعِبِّ مَعَ، فِي هَذَا، مَمْلَكَةٍ مُتَّجِدٍ، فِي سَنَةٍ، مِنْ عُمُرٍ، كُلٌّ مِنْ، تَحْتِ سَنٍّ، سَنٍّ ثَامِنٍ، حَسَبَ إِحْصَاءِ، أَمَكْنُ أَنْ، كَثِيرٌ مِنْ، مِنْ أَمْرِيكِيٍّ، قَدَّرَ عَدَدٌ، عَلَى رَعْمٍ، فِي حِينٍ، عَلَى أَنْ، بَ نَسَمَةٌ، سَاكِنِ بَ، إِلَّا أَنْ، فِي وَقْتٍ، نِسْبَةٌ بَيِّنٌ، عَنِ طَرِيقِ</p>
t-statistics	<p>ولاية مُتَّجِد، وُلِدَ فِي، فِي وِلَايَةٍ، بَلَغَ عَدَدٌ، نِسْبَةٌ مِنْ، عَدَدٌ سَاكِنٍ، عَدِيدٌ مِنْ، لَاعِبِ كُرَّةٍ، ثَامِنٌ عَشْرَةَ، مِنْ قَبْلِ، كُرَّةٌ قَدَّمَ، أَمْرِيكِيٌّ، وُلِدَ، هُوَ لَاعِبٌ، نَسَمَةٌ حَسَبٌ، تَوَقَّى فِي، فِي كُرَّةٌ قَدَّمَ، مِنْ خِلَالِ، أَكْثَرُ مِنْ، حَسَبَ تَعْدَادِ، إِضَافَةٌ إِلَى، لِعِبِّ مَعَ، مَمْلَكَةٍ مُتَّجِدٍ، وَقَعَ فِي، رَعْمٌ مِنْ، تَحْتِ سَنٍّ، سَنٍّ ثَامِنٍ، فِي مَمْلَكَةٍ، حَسَبَ إِحْصَاءِ، مِنْ أَصْلٍ، قَدَّرَ عَدَدٌ، مِنْ عُمُرٍ، بَ نَسَمَةٌ، أَمَكْنُ أَنْ، عَلَى رَعْمٍ، سَاكِنِ بَ، كَثِيرٌ مِنْ، بَيِّنٌ خَامِسٌ، مُتَوَسِّطٌ حَجْمٌ، إِلَّا أَنْ، عَنِ طَرِيقِ، فِي مَنطِقَةٍ، أُسْرَةٌ أُسْرَةٌ، عَدَدٌ أُسْرَةٌ، حَصَلَ عَلَى، نِسْبَةٌ بَيِّنٌ، فِي مَدِينَةٍ، كُلٌّ مِنْ، مِنْ أَجْلِ، فِي حِينٍ، ذُونَ وَجُودِ</p>
Chi-square	<p>ءاد ءاس، ءاس ءيعفوء، ءوعلى ءوبراهيم، ءبحتاجان ءاد، ءيسمك ءبحتاجان، ءيسوان ءورسار، ءينا ءيسمك، ءبى بيجلو، ءجيكالين بلبلو، ءرابينوزيد السيانيدين، ءسائش وپگاه، ءقاي محمدمهدى، ءموى ودرشتى، ءنودية يذوب، ءهنكرانى ومانى، ءگاهى وءود، ءيد همى، ءيزو بوتيل، ءبين كءوردارى، ءبريزى بريكدان، ءبوسعيد ابوالءخير، ءرزوكان ءروزگان، ءربيه الياف، ءزهى ءهود، ءسجايا جيجاىل، ءسقفية سكنية، ءشقى ءهوده، ءعضاؤها ءنفسهم، ءفرادها ءنهم، ءكاديمي للباحءين، ءقابهم وملابسهم، ءلمانيا والسويد، ءنفسهم مافانبو، ءنهم يهود، ءنى اءطيقك، ءوبانىءشاد بريءادار اءياكا، ءوتورينو رسپيگي، ءولية للبيانات، ءشعبى ولا رسميا، ءغلط اءعلمنا، ءناييس لاورينءون، ءكراد فيلية، ءكسيد الكربون، ءكسيروس وءكسيوسيرسا، ءكودي نلءخير، ءولء مين، ءياىءى قرويات، ءي ءبربربي، ءب رهم، ءسءراءىءية وءقءصاءية</p>

Figure A.7: The top 50 collocations from Arabic Wikipedia. (Cheevaprawat-domrong et al., 2022)





จุฬาลงกรณ์มหาวิทยาลัย  
**CHULALONGKORN UNIVERSITY**

## Appendix II

### TOPIC KEYS

EN-NYTimes

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- like one city go street around park long get even day look build come room avenue know home new call</li> <li>- say police two people kill man today charge years three yesterday last former city fire los find officials angeles woman</li> <li>- new york marry son yesterday daughter john church perform rev ceremony robert james university david officiate late richard die paul</li> <li>- first win world last two time open year years television week second women three ago race one lead take show</li> <li>- today united say states government washington american war officials military country president april force south soviet union nations minister march</li> <li>- game season last team national league first night play coach san score one second two run leave players start football</li> <li>- new work one music art book film show theater american play make like even write time york museum dance life</li> <li>- appear would provide could use make many information stand paper new people photograph find alone years may caption item drug</li> <li>- say company million percent yesterday year corporation would share billion market price stock last new bank sell rise offer business</li> <li>- new washington state president city york today school house court would federal public campaign say vote national article bill republican</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- say yesterday police today former charge man case two kill drug find judge arrest die authorities woman court shoot federal</li> <li>- like one even look come make go call new get little use know home place house food room old day</li> <li>- show work theater program dance art <b>new_york</b> music play even american new street open museum present two make hall concert</li> <li>- <b>new_york</b> yesterday marry daughter son rev perform ceremony appear john paper photograph robert caption item david <b>information_provide stand_alone</b> james officiate</li> <li>- today say <b>united_states</b> washington government american april president country may march july meet officials military <b>officials_say</b> moscow <b>united_nations</b> group report</li> <li>- say company yesterday million corporation percent share market business billion bank report price trade offer plan stock group base sell</li> <li>- one city people build two first near water town start take small begin strike home three last drive day early</li> <li>- washington today state say campaign city new vote program would plan may <b>new_york_city</b> march bill house school federal mayor issue</li> <li>- one book life time like seem write world live love think see list come new whose get read know much</li> <li>- team yesterday game victory play win today lead tonight players coach season ap yankees may race <b>last_night</b> last go beat</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- city build town near water house leave park ago home small outside one morning streets travel set day air still</li> <li>- yesterday victory win team lead tonight east play coach players game today ap yankees sport race beat giants season open</li> <li>- today say appear <b>united_states</b> paper american photograph government april caption item <b>information_provide stand_alone</b> country president march military may <b>officials_say</b> july</li> <li>- say company yesterday million percent corporation share market billion business price bank rise stock trade buy offer report group sell</li> <li>- say today yesterday police charge former man drug case federal judge kill arrest authorities court find trial <b>los_angeles</b> hospital friday</li> <li>- washington today state campaign say vote plan new nation march president mayor bill would house call may union program federal</li> <li>- like seem even get come look much make might one go something time know call always think little see good</li> <li>- life article book editor page write list woman whose live years family name read <b>new_york</b> america love story die world</li> <li>- yesterday marry daughter son <b>new_york</b> perform rev ceremony john robert late david officiate james rabbi thomas michael even richard saturday</li> <li>- theater show program work dance play street american art even music museum night sunday open present hall concert Broadway ballet</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- one like make time go get come many seem much even look know could new would think find way might</li> <li>- say company million percent yesterday year would corporation share billion market price stock bank last sell business new offer service</li> <li>- book life television last article week write editor years news two list family show page die time woman story whose</li> <li>- appear music work art information show theater new stand provide <b>new_york</b> street play paper photograph night american alone hall open</li> <li>- washington state today president city would house say federal school public new campaign <b>new_york</b> plan may national vote year bill</li> <li>- city build park one street like room home long around day summer take back come new water small town leave</li> <li>- <b>new_york</b> yesterday marry son daughter church perform rev john ceremony university david robert late officiate michael announce james william even</li> <li>- today say government <b>united_states</b> american washington military officials force country march president war may april united july nations would international</li> <li>- game team last win first season second two national league night play yesterday lead victory time run coach players score</li> <li>- say police two people kill charge today man former yesterday three last years city find drug one fire federal officer</li> </ul>

Figure B.1: The topic keys from the New York Times with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

## EN-SOTU

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- states united government treaty american claim two mexico make last relations citizens governments countries great minister convention receive britain subject</li> <li>- people us one nation country men american know america say come every time let go live work life great good</li> <li>- year amount increase treasury fiscal pay public last expenditures bank revenue estimate present government debt end per sum money june</li> <li>- world nations must peace war free america new continue freedom security people nation economic force international progress us defense strength</li> <li>- new program work need tax federal help job years health make must year budget education school care economy congress million</li> <li>- congress make department service report upon attention present recommend public subject may consideration legislation commission session necessary time provision measure</li> <li>- war land force make navy army military ship officer service naval indian men line number territory indians vessels new coast</li> <li>- states government law power state right would laws constitution united act upon congress shall court case people may without one</li> <li>- business government labor national increase must trade price would country need market production great industry many system foreign large make</li> <li>- country may interest would great upon people government every power peace policy time condition without war good us present public</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- must people great government business men nation one need labor action country work good way matter man would condition deal</li> <li>- country peace interest people nations government may great us every policy power citizens upon nation war national time good cause</li> <li>- states power law laws congress shall government constitution right would may state upon act case subject without people duty authority</li> <li>- program must need new economy job federal congress work budget help tax administration education economic plan provide increase energy continue</li> <li>- trade country commerce american great tariff foreign duties increase upon commercial manufacture market produce value industry price export advantage would</li> <li>- <b>united_states</b> government treaty claim last mexico governments convention th question make subject upon minister receive <b>great_britain</b> spain two international republic</li> <li>- amount year public expenditures treasury government increase bank present estimate pay debt <b>fiscal_year</b> revenue sum make interest receipt money note</li> <li>- america world us people nation know freedom americans must tonight every let american nations free today new peace future live</li> <li>- congress make report recommend present subject department service consideration work land attention upon commission legislation message necessary system last important</li> <li>- war force military service navy time army vessels officer require peace ship number arm one defense necessary enemy country troop</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- country business labor trade price great increase tariff market industry foreign benefit american manufacture condition farm system produce value agriculture</li> <li>- nation people america world us must peace freedom great country free know stand men american fight live future life every</li> <li>- law states power laws constitution government right shall upon congress people case act authority exercise would without question prevent purpose</li> <li>- <b>united_states</b> treaty government claim mexico governments th convention subject citizens make receive minister relations question upon spain <b>great_britain</b> act republic</li> <li>- job americans america help tax education budget children must work reform congress economy families every need ask program tonight let</li> <li>- must program continue economic need new world nations administration security nation national defense development progress policy strengthen support action cooperation</li> <li>- amount year expenditures treasury government increase present bank public estimate debt pay revenue sum <b>fiscal_year</b> money receipt interest expense millions</li> <li>- congress report recommend subject present make legislation consideration department upon commission attention session service necessary provision message system last view</li> <li>- country interest government upon peace may great every without citizens people nations policy war power us would condition national good</li> <li>- land service navy war territory indians army vessels military force ship make line indian necessary construction islands</li> <li><b>united_states</b> place carry</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- world people us nation peace nations america must free freedom war american one time force men great come together live</li> <li>- year amount increase treasury public bank fiscal last expenditures pay estimate debt government present end revenue money years june sum</li> <li>- country may interest government upon would great power people every policy peace without time citizens us condition good cause foreign</li> <li>- land war force navy army military service make ship line officer naval indian men number territory vessels indians one two</li> <li>- congress make department report service public subject attention recommend may present upon consideration necessary provision session law legislation commission time</li> <li>- program federal new national need must economic government continue provide development increase energy administration resources economy private policy develop budget</li> <li>- <b>united_states</b> government treaty american claim two make countries mexico last relations citizens governments minister convention subject th britain receive great</li> <li>- would country business labor great trade system market price public commerce increase foreign tariff upon condition national large result must</li> <li>- government power states state right law constitution would congress people shall laws <b>united_states</b> act court upon case one union exercise</li> <li>- work people make job americans every america tax american new help children care health must need one years year go</li> </ul>

Figure B.2: The topic keys from the United States State of the Union Addresses with different retokenization measures. (Cheevaprawatdomrong

## EN-Yelps

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- time go get take service work would call make need say back tell give ask experience never staff car care</li> <li>- place sandwich good coffee breakfast get love great lunch go try always like time make ice food egg friendly one</li> <li>- order get food go us come time wait would table back take service say ask one minutes give could sit</li> <li>- great place bar good food drink go beer service happy nice night love hour time atmosphere really patio wine always</li> <li>- food good place chicken roll sushi dish eat like order try restaurant rice great service lunch soup go taste menu</li> <li>- salad order taste menu cheese bread delicious try dish flavor meal restaurant dessert make also sauce good dinner side serve</li> <li>- like get place know go say think one people really want make look see time could star review would way</li> <li>- store find price shop like go buy one place get love great always look selection need location really lot also</li> <li>- good pizza food fry place burger get like cheese order eat chicken sauce go mexican try taste really chip best</li> <li>- room park get phoenix area one nice stay lot see hotel pool free walk also great around time day go</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- store shop price find one buy get like selection go look need location stuff items love place use always see</li> <li>- room nice park phoenix area stay little hotel pool seat outside great free also clean beautiful inside scottsdale small place</li> <li>- food order service table us wait come restaurant get sushi server go time ask dinner experience sit eat good seat</li> <li>- get work go call time need car take staff service experience would give even tell come use dog new wait</li> <li>- flavor taste delicious try mexican cheese chip salsa bean menu chocolate like good dessert tacos make sweet eat cake food</li> <li>- place bar great drink beer happy good go like hour get night pretty wine cool fun crowd love music friends</li> <li>- place great food good service always love friendly staff best go coffee price try amaze awesome time location atmosphere nice</li> <li>- like go say place get one know think bad ask people want review even give look tell see star guy</li> <li>- sandwich fry good burger order cheese salad get breakfast like place bread chicken eat lunch side food come egg try</li> <li>- pizza food good chicken place sauce order dish like eat taste soup try rice lunch restaurant spicy thai menu fresh</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- great food place good service always love friendly beer staff price nice atmosphere lunch awesome happy go get bar drink</li> <li>- food chicken good dish sauce eat order place rice like taste restaurant soup spicy meat lunch try thai roll beef</li> <li>- bar room nice park place area seat drink fun phoenix crowd stay walk great hotel pool get cool night outside</li> <li>- pizza sandwich salad bread cheese flavor good fresh like try sauce taste chocolate delicious order get sweet eat love little</li> <li>- store shop price find buy location selection look items like always go need stuff love great mall everything one see</li> <li>- restaurant menu wine dinner sushi food order service great drink table happy meal bar experience enjoy server hour us dish</li> <li>- get dog car go need staff work recommend clean experience call take friendly would give service care even find tell</li> <li>- place get coffee like love know go drink oh think good say make try review ca yes thing let well</li> <li>- fry food good burger breakfast cheese order eat get mexican place chip salsa like chicken egg tacos taste bean try</li> <li>- wait get ask order go food table say come us service minutes bad place tell sit even experience think leave</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- menu wine restaurant dinner great salad order dish delicious meal dessert also us make enjoy bite taste well good cheese</li> <li>- good coffee breakfast place like get mexican ice food try flavor cream chip make salsa egg taste tacos love one</li> <li>- bar place drink great beer park good like night nice go area people music fun see play game lot pretty</li> <li>- get take time go would work room need call car staff service make clean dog day use great stay hotel</li> <li>- food good place chicken roll like sushi order dish eat rice restaurant sauce taste try lunch soup really spicy come</li> <li>- like know one say think go get make place would review star time people give want see could look something</li> <li>- place great food good service pizza love always go time friendly staff best price get really try happy atmosphere lunch</li> <li>- go get order us come time would wait food back take say ask service table minutes tell one could give</li> <li>- good fry sandwich like get burger cheese order chicken sauce salad go eat bread side try really place come lunch</li> <li>- store find shop price go like one get buy love look always selection location need great place items lot stuff</li> </ul>

Figure B.3: The topic keys from the Yelp Dataset with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

## DE-10kGNAD

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- Die Wien SPÖ sei ÖVP Der FPÖ Wiener Partei Das Grünen Wahl ORF Hofer Prozent Er Auch Van Bürgermeister Kandidaten</li> <li>- Die Der Bis of New The Im Wien Welt Jahren In Jahre Ein ORF Grad Film Uhr Geschichte Mit Am</li> <li>- Die Der worden sei Polizei wegen Das Er sagte Ein Mann In zwei laut gewesen Nach Frau Jahre seien Staatsanwaltschaft</li> <li>- Die Flüchtlinge sagte Österreich sei Deutschland EU Das Regierung Menschen In Griechenland Der Flüchtlingen Europa Es Land Grenze seien Wir</li> <li>- Die Das Apple neue Nutzer Google Der gibt neuen Facebook Windows Mit Microsoft Unternehmen So Internet Daten Hersteller Für könnte</li> <li>- Die Der worden Menschen sagte USA Regierung Russland IS Syrien In seien Präsident Angaben sei getötet Nach Staat zwei vergangenen</li> <li>- Sie STANDARD Das Es Ich Und Wir gibt Aber sagt Wenn gut geht ganz Er Was Man Wie Zeit Die</li> <li>- Die Wien Das Menschen In Diese Forscher sagt gibt Jahren Kinder sei Österreich Studie Eine Universität Bei So könnten Sie</li> <li>- Der Die ersten zwei In Spiel drei Platz Salzburg Sieg FC zweiten Saison sagte Minute Punkte Nach vier Austria Trainer</li> <li>- Prozent Euro Die Millionen Der Jahr rund Milliarden Das Österreich Jahren Unternehmen Dollar Im Wien In zwei sei Geld drei</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- Die Forscher Wien Der In Wissenschaftler Menschen Grad Das Studie Erde Tiere Uni Bei Diese konnten zwei Ein Experten sagt</li> <li>- Prozent Die Euro Der Jahr Österreich Unternehmen Das sei Wien <b>Millionen_Euro</b> Geld laut Millionen seien Dollar Bei Auch <b>Milliarden_Euro</b> Zahlen</li> <li>- Die Der sagte Spiel Wir Ich Sieg Minute Salzburg Sonntag Rapid Trainer Austria Samstag Es Nach Mannschaft ersten APA gut</li> <li>- Die sagte Regierung Der sei USA Das Präsident wegen Russland EU Land Auch Iran Entscheidung Dienstag erklärte Parlament Donnerstag worden</li> <li>- Sie STANDARD Das Ich Und Es sagt Wir Aber Wenn gibt Was gut ganz Wie geht Man Da Menschen Die</li> <li>- Die Apple Nutzer Google Das Facebook neuen neue gibt Der Microsoft Daten Mit Für Windows Auch Unternehmen Hersteller So Smartphone</li> <li>- Flüchtlinge Österreich Die sagte Menschen sei Flüchtlingen Deutschland Grenze Türkei EU seien Europa In Das Montag Auch Land Asylwerber Wir</li> <li>- Wien Die SPÖ sei ÖVP Wiener FPÖ Der Grünen Das Hofer Partei In Auch Für ORF Strache laut Österreich Häupl</li> <li>- Der Bis Die ORF Welt USA steht The In Geschichte Wien Ein Mit Trump Als Film Das Er Bühne Verfügung</li> <li>- Die Der Polizei sagte sei seien worden IS Ein Syrien Angaben Mann Bei In Er wegen zwei laut berichtete Opfer</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- Bis Der ORF Welt In The Wien Die USA Mit Geschichte Als Ein Film Wiener Musik Das Bühne <b>New_York</b> Buch</li> <li>- Der Die Partei Hofer sei ORF FPÖ Er Wahl Kandidaten Entscheidung Das <b>Van_Bellen</b> Parteien Strache könnte Wien wegen Fall Stimmen</li> <li>- Die Forscher Wien Wissenschaftler Grad Menschen Studie In Erde Das Der Tiere Diese Uni konnten sagt Wasser Für Ein Eine</li> <li>- sagte Die Regierung Türkei Syrien Der Griechenland Russland IS USA Land Montag sei EU seien Präsident Brüssel Europa Sonntag vergangenen</li> <li>- Der Die sagte Spiel Ich Sieg Wir Minute Rapid Trainer Salzburg Samstag Austria Sonntag APA Mannschaft Nach Tore Für ersten</li> <li>- Prozent Die Euro Der Österreich Unternehmen Jahr Das Geld Millionen laut <b>Millionen_Euro</b> sei Auch seien Zahlen Für Dollar Banken vergangenen</li> <li>- Die Apple Nutzer Google neue Facebook Das gibt Microsoft neuen Mit Windows Daten Auch Der Hersteller Unternehmen Für könnte So</li> <li>- Die Polizei Der sagte worden sei Ein wegen seien Mann laut Bei Er Opfer berichtete gewesen zwei Menschen Polizisten In</li> <li>- Flüchtlinge sei Österreich Wien SPÖ ÖVP Die sagte Flüchtlingen seien Auch Das Wiener müsse Faymann FPÖ Menschen In Für Regierung</li> <li>- STANDARD Sie Das Ich Und Es Wir sagt Aber Wenn gibt Was gut ganz Wie geht Man Da Menschen Die</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- Die Menschen sagte Der Türkei worden Syrien Regierung Russland seien USA IS Angaben In sei Land getötet Nach Staat Präsident</li> <li>- Die Der sei sagte Regierung Das EU Er wegen Deutschland Griechenland worden In Entscheidung Parlament Partei deutsche Zeitung Auch Merkel</li> <li>- Prozent Euro Die Millionen Der Jahr Österreich rund Das Milliarden Jahren Wien Unternehmen Dollar In Im sei drei zwei laut</li> <li>- Die Wien Flüchtlinge Österreich sei SPÖ ÖVP FPÖ Das Wiener In Der sagte Flüchtlingen Grünen Auch Faymann seien Für Hofer</li> <li>- Die Der Polizei worden sei wegen Mann Ein sagte Er zwei In Das drei gewesen Am Jahre Frau laut Nach</li> <li>- Der Die ersten zwei drei Spiel In Platz Salzburg FC zweiten Sieg Saison Wir sagte vier Minute Nach Austria Trainer</li> <li>- Die Das Apple neue Nutzer gibt Google Der neuen Facebook Windows Mit Microsoft Unternehmen Auch Daten So Internet Hersteller Für</li> <li>- Sie Das Es STANDARD Ich Und gibt Wir sagt Aber Die Wenn geht gut ganz Was In Man Wie Menschen</li> <li>- Der Die Bis ORF New The Wien Das of Welt Geschichte York In Film USA the Jahre Ein Als Im</li> <li>- Die Forscher Der Wien Das Universität Jahren Menschen In Studie Wissenschaftler steht Verfügung Tiere Erde Diese Im Gründen rund Uni</li> </ul>

Figure B.4: The topic keys from the the Ten Thousand German News Articles Dataset with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

CN-Chinanews

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- 场比赛 赛季 世界杯 时间 昨天 两 轮 北京 联赛 球队 球员 队 新 总 已经 次 主场 支 巴西</li> <li>- 元 增长 中国 数据 显示 同比 发布 经济 价格 \.亿 国家 今年 公布 全 月份 电 国 去年 美元 记者</li> <li>- 中国 世界 记者 电 国际 届 北京 位 中新网 冠军 首 比赛 次 举行 昨天 第一 名 韩国 体育 上海</li> <li>- 岁 中新网 节目 网友 媒体 消息 近日 两 没 网络 天 没有 生活 不少 表示 已经 日前 卫视 接受 台湾</li> <li>- 电影 记者 中国 历史 — 奖 部 作品 位 艺术 北京 摄 文物 影片 导演 文学 出版 著名 票房 博物馆</li> <li>- 中国 经济 发展 表示 会议 国际 问题 国 改革 新 专家 国家 次 社会 政府 政策 组织 工作 重要 关注</li> <li>- 公司 市场 家 元 基金 集团 企业 银行 上市 两 投资 行业 产品 新 出现 人士 成为 股 周 目前</li> <li>- 上海 记者 工作 部门 电 获悉 市 全 国家 上海市 新闻 网站 服务 管理 今年 安全 发布 信息 北京 三峡</li> <li>- 报道 电 中新网 名 美国 香港 台湾 日本 媒体 英国 消息 发生 组织 表示 两 调查 事件 发现 宣布 人员</li> <li>- 电 两 记者 岸 中新网 文化 中国 举行 台湾 中新社 北京 上海 届 活动 新 大学 合作 论坛 交流 海峡</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- 今年 电影 观众 — 部 近日 导演 片 位 场 奖 影片 晚 成为 明 星 中国 昨日 剧 版 网络</li> <li>- 中国 昨天 比赛 记者 赛季 北京_时间 场 世界杯 球队 最终 上海 昨晚 球迷 球队 结束 本报 讯 强 冠军 男篮</li> <li>- 记者 摄 — 天 历史 旅游 图 发现 新华社 今年 广州 南京 专家 旅客 游客 文物 前 市民 新 号</li> <li>- 中国 表示 发展 专家 问题 新 经济 国 改革 未来 政策 关注 会议 合作 国家 重要 成为 推动 峰会 国际</li> <li>- 岁 两 网友 没有 消息 种 里 近日 没 很多 位 照片 日前 不少 想 前 媒体 已经 表示 现在</li> <li>- 上海 记者 工作 电 获悉 部门 上海市 企业 市 中新网 近日 日 前 情况 国家 实施 北京 昨天 服务 目前 今年</li> <li>- 价格 中国 今年 发布 增长 月份 电 指数 去年 上涨 下降 中新网 经济 公布 同比 数据 记者 消费 期 全 国</li> <li>- 电 中新网 报道 香港 台湾 名 日本 美国 消息 韩国 表示 英国 发生 警方 外媒 中央社 东京 遭 俄罗斯 今 日</li> <li>- 电 记者 中新网 北京 中新社 两_岸 中国 举行 上海 台湾 新 活动 国际 论坛 社 文化 开幕 世界 海峡_两_岸 大陆</li> <li>- 市场 元 家 基金 公司 银行 股 投资 行业 投资者 今年 资金 产 品 成为 目前 周 交易 发行 中国 集团</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- 昨天 赛季 比赛 世界杯 北京_时间 中国 球员 场 记者 队 昨晚 本报 讯 球迷 球队 强 已经 主场 最终 男篮 对手</li> <li>- 电 记者 中新网 中新社 北京 两_岸 上海 举行 文化 新 论坛 中国 台湾 社 旅游 开幕 合作 活动 — 海峡_两_岸</li> <li>- 市场 元 价格 今年 增长 月份 发布 中国 去年 指数 基金 上涨 期 下降 同比 数据 公布 消费 增速 百分点</li> <li>- 岁 消息 两 不少 没有 网友 日前 里 近日 没 位 很多 出 最近 中新网 想 表示 照片 张 昨日</li> <li>- 上海 记者 企业 工作 家 近日 日前 实施 上海市 获悉 服务 市 部门 国家 公司 昨天 信息 标准 网站 今年</li> <li>- 中国 经济 表示 专家 发展 关注 改革 未来 政策 问题 新 国 认 为 会议 成为 投资 重要 积极 面临 峰会</li> <li>- 今年 观众 电影 — 近日 片 导演 作品 昨日 影片 成为 部 亮相 拍 卖 网络 位 剧 演出 品牌 明星</li> <li>- 中新网 电 报道 香港 台湾 名 日本 消息 韩国 外媒 发生 警方 中央社 英国 美国 东京 遭 事件 表示 文 汇报</li> <li>- 电 中新网 记者 中国 北京 世界 国际 上海 举行 广州 中新社 新 成都 体育 帷幕 拉 开 站 赛事 晚 今年</li> <li>- 记者 摄 天 价格 新华社 — 前 三峡 图 专家 今年 历史 南京 发现 新 获悉 了解 天气 市民 本报</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- 电 两 中国 记者 岸 北京 举行 中新网 中新社 上海 届 台湾 国 际 合作 大学 新 文化 活动 论坛 交流</li> <li>- 中国 经济 发展 表示 新 改革 企业 社会 工作 国 政府 国家 中 央 政策 建设 问题 会议 金融 全 上海</li> <li>- 比赛 中国 场 赛季 世界 昨天 北京 冠军 世界杯 轮 记者 时 间 决 赛 两 次 联赛 球队 球员 总 队</li> <li>- 电影 部 奖 观众 节目 位 导演 中国 首 北京 今年 影片 电 视 文学 昨日 近日 卫视 票房 片 著名</li> <li>- 种 没有 已经 关注 不少 成为 出 天 里 现在 出现 生活 两 没 表示 前 一直 认为 很多 可能</li> <li>- 中新网 电 香港 岁 台湾 报道 名 消息 两 三峡 表示 韩国 成员 前 警方 照片 日前 新 男子 中央社</li> <li>- 记者 中国 文化 历史 电 中新网 — 摄 专家 艺术 南京 博物馆 文 物 位 次 件 图 地区 成都 保护</li> <li>- 电 报道 美国 中新网 日本 国际 世界 中国 组织 名 全球 国家 英国 表示 宣布 时间 媒体 次 东京 国</li> <li>- 公司 记者 上海 家 元 集团 网站 获悉 部门 工作 昨天 有 限 新 闻 市 消息 安全 企业 人员 目前 全</li> <li>- 元 增长 市场 中国 价格 数据 显示 经济 同比 发布 \.亿 今年 公布 月份 国家 基金 美元 全 投资 上涨</li> </ul>

Figure B.5: The topic keys from the Chinanews with different retokenization measures. (Cheevaprawatdomrong et al., 2022)



## CN-Dianping

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- 家吃次店喜欢现在以前没很多经常感觉不错每次一直东西觉得排队好吃真的开</li> <li>- 买东西价格比较便宜家很多逛贵喜欢超市里面活动经常系店挺感觉卖质量</li> <li>- 吃味道元不错好吃份烤饭两套餐新鲜没有海鲜种东西太觉得少次牛肉</li> <li>- 没没有服务员知道想太服务次钱差店家块实在态度两真以后最后问</li> <li>- 不错里面环境地方挺比较感觉太位置楼朋友次很多晚上方便房间玩舒服酒店电影</li> <li>- 不错比较环境菜味道服务感觉价格挺贵家口味朋友算适合地方有点觉得特别高</li> <li>- 次非常天没想电话公司一直种拍真的医院服务没有两态度之前时间感觉一下</li> <li>- 地方很多种路里面走上海非常公园看到里感觉没有现在北京两车街真的门口</li> <li>- 吃味道好吃菜汤不错鱼辣肉碗喜欢有点两里面感觉太面鸡挺锅</li> <li>- 喜欢吃家味道好吃喝买不错蛋糕杯感觉面包种里面咖啡甜太店觉得奶茶</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- 喝环境咖啡杯地方茶喜欢系下午朋友感觉舒服位置太坐店真家坐在饮料</li> <li>- 里面地方很多不错比较玩方便挺环境感觉房间公园走电影元住蛮坐晚上酒店</li> <li>- 没服务员没有服务态度差想问最后知道以后太后来钱两居然电话换天元</li> <li>- 吃味道好吃家不错面喜欢饭碗感觉汤觉得没烤次挺店份牛肉太</li> <li>- 吃味道好吃喜欢不错家蛋糕里面面包买感觉甜觉得套餐元新鲜没有种太口感</li> <li>- 不错环境比较感觉服务价格菜味道挺贵朋友口味家地方算吃饭觉得总体特色性价比</li> <li>- 种觉得没里感觉知道真的其实看到一下朋友没有非常真想一定应该发现点评老板</li> <li>- 买东西比较很多价格家喜欢便宜里面逛超市贵活动经常挺衣服卖店不错感觉</li> <li>- 吃味道菜好吃鱼不错汤辣肉太锅有点里面两挺虾比较香炒</li> <li>- 家次现在以前没很多吃经常后来排队家_店太开一直每次没有感觉知道觉得记得</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- 非常看到朋友觉得想没真的其实点评种一定天一下拍老公真是已经最后一直真</li> <li>- 没服务员没有服务差态度知道太想钱问以后东西后来店吃实在家居然真</li> <li>- 吃味道好吃汤菜不错鱼辣家肉元碗有点里面锅面喜欢香油牛肉</li> <li>- 系好多医院次医生上海真知道很多路非常间食其实元算度真的好好老</li> <li>- 地方环境不错里面感觉很多挺楼舒服喜欢里走房间坐酒店蛮晚上比较位置住</li> <li>- 吃家味道喜欢好吃喝买不错蛋糕感觉杯面包太甜咖啡种奶茶觉得次里面</li> <li>- 买东西比较家价格很多喜欢贵逛便宜里面店经常超市不错挺感觉卖衣服活动</li> <li>- 不错比较环境菜味道服务价格感觉挺口味家贵吃朋友算适合太没吃饭地方</li> <li>- 吃味道不错好吃饭份家烤喜欢觉得套餐感觉元没比较次中午新鲜店很多</li> <li>- 现在以前次很多元比较好像感觉经常便宜地方办没卡玩一直没有挺时间家</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- 吃味道好吃喜欢家汤不错面份两辣碗肉感觉元挺烧烤觉得牛肉</li> <li>- 吃喜欢家味道好吃喝蛋糕杯买面包咖啡种甜不错店奶茶里面太茶元</li> <li>- 不错环境比较感觉挺服务地方里面朋友位置楼装修方便有点适合舒服太价格晚上算</li> <li>- 家次现在以前没很多店没有经常一直开后来排队好像每次记得朋友蛮喜欢知道</li> <li>- 买东西比较价格家便宜店很多贵喜欢逛超市活动里面经常系不错卖质量感觉</li> <li>- 很多地方里面种公园走非常里感觉看到路两上海街环境看看条北京书其实</li> <li>- 菜吃味道不错鱼好吃元比较汤有点没有虾推荐太鸡海鲜新鲜肉烧感觉</li> <li>- 吃不错味道比较家价格环境感觉次贵店觉得挺服务口味菜喜欢东西太好吃</li> <li>- 没没有服务员服务想太知道差两吃实在家钱次店态度块居然最后以后</li> <li>- 次天非常想时间电话公司医院没一下没有张知道一直拍元真的态度之前第一</li> </ul>

Figure B.6: The topic keys from the Dianping with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

CN-Douban

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- 美队 钢铁侠 快 蜘蛛 队长 心疼 太 女巫 帅 银 黑 美队 喜欢 寡妇 蜘蛛侠 绿巨人 冬兵 可爱 死</li> <li>- 电影 觉得 部 没有 真的 太 看完 想 没 高 期待 片子 种 知道 很多 好看 喜欢 失望 其实 看到</li> <li>- 太 故事 剧情 人物 没有 有点 感觉 情节 觉得 不错 没 很多 感情 喜欢 角色 部分 实在 够 讲 其实</li> <li>- 部 剧情 好看 没有 特效 感觉 第一 场面 电影 不错 有点 打 斗 太 没 精彩 漫威 爆米花 挺 大片 种</li> <li>- 英雄 超级 漫威 电影 美国 蜘蛛侠 队长 复仇者 系列 联盟 大战 内战 部 精彩 新 两 妇联 反派 角色 里</li> <li>- 动画 画面 故事 国产 中国 剧情 画风 不错 宫崎骏 电影 国漫 制作 动漫 支持 美 部 希望 值得 风 音乐</li> <li>- 没 次 两 睡着 看的 电影院 想 看到 看完 电影 彩蛋 前 第一 半 好看 遍 真 的 一直 小时 太</li> <li>- 爱 爱情 喜欢 鱼 最后 女主 没有 世界 男 牺牲 湫 条 感动 生命 种 事 故事 爷爷 句 大鱼</li> <li>- 没 想 真 种 编剧 脑 句 逼 钱 真是 知道 坑 死 懂 里 简直 片 全 完全 出</li> <li>- 画面 星 剧情 尴尬 台词 故事 太 女主 配音 音乐 真的 颗 不 错 两 配乐 美 男 三星 胎 观</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- 爱 女主 爱情 喜欢 男 鱼 最后 种 胎 牺牲 感动 湫 备 生命 世界 没有 故事 爷爷 死 椿</li> <li>- 好看 剧情 感觉 有点 没 不错 没有 太 挺 觉得 喜欢 <b>第一_部</b> 特效 失望 场面 打 斗 精彩 其实 差 蛮</li> <li>- 觉得 电影 没有 想 真的 没 看完 片子 知道 种 部 <b>部_电影</b> 黑 期待 情怀 看到 豆瓣 评论 懂 失望</li> <li>- 美 队 蜘蛛侠 蜘蛛 钢铁侠 心疼 美队 队长 蚁人 太 冬兵 可爱 <b>美国_队长</b> 帅 黑豹 最后 虐 冬 托尼 喜欢</li> <li>- 太 人物 故事 剧情 没有 情节 感觉 感情 逻辑 导演 角色 很多 讲 有点 够 世界观 简单 略 细节 部分</li> <li>- 漫威 英雄 部 电影 <b>超级_英雄</b> 种 片 特效 爆米花 大片 打 斗 系列 精彩 妇联 场面 反派 美国 疲劳 拍 已经</li> <li>- 画面 动画 国产 中国 剧情 不错 故事 画风 宫崎骏 国漫 动漫 很多 支持 值得 美 希望 风 大鱼 真的 确实</li> <li>- 快 女巫 银 死 太 奥 创 <b>黑_寡妇</b> 绿巨人 喜欢 最后 幻 视 没有 寡姐 猩 红 眼 居然 没 彩蛋 帅 红</li> <li>- 画面 剧情 星 台词 尴尬 太 故事 配音 音乐 配乐 女主 真的 两 三星 不 错 观 <b>颗_星</b> 美 半 烂</li> <li>- 没 电影 想 睡着 看完 电影院 看的 看到 遍 感觉 两 影院 彩蛋 <b>第一_次</b> 哭 钱 分钟 一起 睡 真的</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- 故事 人物 太 剧情 没有 情节 很多 感觉 电影 细节 讲 角色 感情 表达 导演 比较 简单 不够 逻辑 整体</li> <li>- 快 女巫 银 太 死 喜欢 奥 创 <b>黑_寡妇</b> 绿巨人 没有 最后 幻 视 彩蛋 猩 红 眼 没 寡姐 钢铁侠 红 妇联</li> <li>- 漫威 英雄 打 斗 特效 场面 电影 部 精彩 <b>超级_英雄</b> 剧情 种 爆米花 大片 片 大战 系列 不错 妇联 过瘾 期待</li> <li>- 美 队 蜘蛛 钢铁侠 心疼 蜘蛛侠 美队 队长 蚁人 帅 <b>美国_队长</b> 冬兵 可爱 黑豹 太 最后 虐 冬 喜欢 钢铁</li> <li>- 好看 没 觉得 感觉 没有 太 有点 剧情 挺 不错 失望 期待 真的 喜欢 其实 <b>第一_部</b> 差 可能 部 后面</li> <li>- 画面 剧情 星 台词 故事 太 尴尬 音乐 配音 配乐 不错 美 真的 女主 三星 很美 <b>颗_星</b> 画风 观 情怀</li> <li>- 想 没 真 编剧 逼 太 出 梦 知道 真是 坑 好 好 种 里 片 钱 心 吃 不要 垃圾</li> <li>- 电影 没 看完 想 睡着 电影院 看的 看到 遍 脑 影院 觉得 真的 次 种 哭 片子 一起 知道 彩蛋</li> <li>- 爱 女主 男 爱情 鱼 喜欢 最后 感动 没有 胎 牺牲 故事 湫 备 生命 世界 爷爷 种 海棠 椿</li> <li>- 动画 国产 中国 画面 电影 觉得 国漫 画风 大鱼 宫崎骏 部 真的 希望 很多 支持 动漫 值得 不错 海棠 故事</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- 爱 女主 男 最后 喜欢 爱情 鱼 没有 感动 胎 牺牲 条 湫 备 生命 死 故事 世界 想 句</li> <li>- 快 黑 喜欢 女巫 银 寡妇 死 太 奥 创 绿巨人 没有 最后 没 幻 视 眼 猩 红 寡姐 复 联 钢铁侠 里</li> <li>- 英雄 电影 超级 漫威 种 部 爆米花 大片 片 美国 联盟 复仇者 系列 里 已经 世界 疲劳 拍 内战 审美</li> <li>- 没 电影 好看 次 部 看完 觉得 想 第一 睡着 看的 电影院 两 感觉 看到 没有 前 真的 遍 挺</li> <li>- 剧情 画面 不错 太 真的 故事 音乐 有点 美 台词 挺 配乐 尴尬 觉得 配音 情节 棒 确实 烂 赞</li> <li>- 故事 画面 中国 讲 剧情 画风 台词 人物 太 情节 宫崎骏 设定 配音 尴尬 风 美 爱情 喜欢 元素 配乐</li> <li>- 星 没 剧情 想 两颗 观 情怀 真 编剧 半 脑 钱 全 片 加 本来 三星 烂 简直</li> <li>- 太 没有 部 剧情 感觉 打 斗 好看 人物 场面 精彩 有点 失望 第一 期待 不错 戏 比较 角色 特效 节奏</li> <li>- 美 队 队长 蜘蛛侠 蜘蛛 钢铁侠 心疼 美队 蚁人 美国 太 冬兵 可爱 帅 好看 真的 黑豹 最后 虐 冬</li> <li>- 电影 动画 国产 部 觉得 没有 国漫 很多 大鱼 希望 支持 真的 看到 动漫 海棠 值得 中国 已经 作品 大圣</li> </ul>

Figure B.7: The topic keys from the Douban with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

JA-JapanNews

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- さん 為る 風 テレビ 話 番組 後 放送 第 監督 ドラマ 時 日 出演 役 メートル 強い 視 聴 月 時間</li> <li>- 為る 日 月 年 居る 映画 成る 公開 曲 ファン イベント 此の 作品 開催 ライブ 達 中 御 写真 登場</li> <li>- 為る 居る 者 事 年 有る 月 つく 氏 成る 万 言う 会社 企業 会 日 的 党 問題</li> <li>- 為る 御 居る 成る 有る 食べる 店 事 円 出来る 中 様 物 時 言う 期間 日 使う 商品 見る</li> <li>- 為る 日 位 日本 年 選手 戦 回 試合 監督 第 居る 野球 チーム 大会 成る リーグ 目 代表 優勝</li> <li>- 為る 居る 事 有る 出来る 成る 性 因る 様 的 可能 開発 車 研究 此の 言う 年 使う 為 技術</li> <li>- 為る 居る 日本 事 人 中国 有る 年 国 成る 的 日 アメリカ 因る 為 世界 政府 者 感染 此の</li> <li>- 為る 居る 言う 事 有る 無い 成る 人 様 思う 良い 其の 見る ね 行く そう たい 御 来る よ</li> <li>- 為る 居る 日 県 市 人 有る 者 月 情報 因る 時事 言う 円 年 区 内容 疑 日本</li> <li>- ユーザー 機能 中 利用 募集 記事 評価 出来る コメント ニュース 投稿 メッセージ 通知 円 各種 レビュー 成る 新規 新しい ドル</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- 為る 言う 人 有る 無い 居る ね 成る 自分 思う 良い 其れ 様 見る 其の 事 そう って 女性 男性</li> <li>- 風 メートル 後 時 為る 強い 天気 波 や や 居る 体 人 攻略 確率 因る 有る 時間_帯 デング 医師 患者</li> <li>- 為る 情報 日 因る 居る さん 有る 御 月 ニュース 為る_居る 町 市 人 見る 仕事 男性 アクセス ランキング 在宅</li> <li>- 日 選手 位 監督 為る 年 チーム 大 日本 巨人 ロッテ 投手 試合 野球 回 戦 御 纏め 今季 スポーツ</li> <li>- 為る 御 此の 使う 電池 アプリ 出来る ボタン 有る 様 食べる モデル 円 見る 商品 言う 此れ 置く 因る 店</li> <li>- さん 為る 日 ドラマ 言う ファン 映画 年 写真 役 月 番組 演 ずる 此の 自身 中 御 人気 姿 氏</li> <li>- 為る イベント ライブ 月_日 アニメ 音楽 曲 達 第_話 会場 此の 御 ゲーム 成る ステージ 等 公演 ツアー 作品 映像</li> <li>- 月 企業 為る 成る 市場 サービス データ 年 円 因る つく 日 ドル 社 技術 氏 中国 関連 調査 今後</li> <li>- ユーザー 中 募集 記事 ニュース 機能 コメント 評価 投稿 メッセージ 通知 各種 レビュー 機能_利用_出来る 新規 成る 新しい わあ ザップ 結核</li> <li>- 氏 為る 日本 日 中国 つく ニュース 関連 国 対する 政府 為る_居る アメリカ 述べる 有る 中 巡る 此れ 無い 成る</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- 此の アプリ 因る サービス 出来る 使う モデル 技術 型 システム 此れ 氏 言う 用 スマート_フォン 製品 画像 動画 データ カメラ</li> <li>- 氏 日本 日 つく 関連 ニュース 中国 国 対する 政府 巡る 向ける アベ 為る 述べる アメリカ 首相 組織 情報 シリア</li> <li>- さん 日 因る 情報 居る 男性 為る 月 ニュース 男 警察 事件 マンション つく 有る 容疑_者 人 元 件 市</li> <li>- 為る 食べる 御 体 方 結核 言う 有る 中 効果 味 物 癌 成る 美味しい 使う 時 菌 健康 出来る</li> <li>- 選手 位 日 監督 チーム 巨人 ロッテ 大 投手 野球 試合 回 日本 纏め スポーツ 今季 年 球 団 点 戦</li> <li>- さん ファン ドラマ 映画 監督 達 ライブ 役 日 人気 第_話 演 ずる 言う アニメ 音楽 曲 作品 等 番組 女優</li> <li>- 企業 円 市場 ドル 中国 月 成る 日 つく 社 比 米 事業 平均 居る 攻略 海外 億_円 日本 大手</li> <li>- 御 為る 電池 此の ボタン 町 店 トウキョウ 情報 下 される 会場 さん イベント 頂く 日 猫 円 月_日 様 市</li> <li>- 言う 無い ね 思う 為る 自分 有る 其れ 良い 人 って さん 居る 見る 様 そう 御 成る 女性 てる</li> <li>- ユーザー 中 募集 記事 ニュース コメント 評価 レビュー 機能 投稿 メッセージ 通知 各種 機能_利用_出来る 新規 風 成る 新しい わあ ザップ</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- 風 為る メートル 食べる 御 後 電池 強い 時 天気 ボタン 波 居る や や 料理 美味しい 味 時間_帯 猫 確率</li> <li>- 為る 有る 体 因る 此の 成る 研究 結核 人 言う 患者 デング 健康 其の 効果 癌 出来る 中 物 熱</li> <li>- 月 企業 成る 年 市場 為る 円 日 中国 ドル 因る 社 サービス 日本 米 つく 会社 比 有る 平均</li> <li>- さん 為る ファン 日 映画 ドラマ 年 達 作品 監督 イベント 月_月_日 成る 言う ライブ 役 此の 人気 等</li> <li>- 氏 日本 為る 日 つく 中国 対する 国 ニュース 関連 政府 アメリカ 述べる 有る 此の 因る 中等 巡る 韓国</li> <li>- 為る 言う 有る 居る 人 無い 成る 様 ね 思う 自分 良い 見る 其れ 其の 事 御 そう 女性 って</li> <li>- 為る 日 情報 因る 居る さん 有る 人 ニュース 町 市 為る_居る 見る 警察 ランキング 月 男性 在宅 アクセス 事件</li> <li>- 御 為る 此の 有る アプリ 使う 因る モデル 様 言う 下 される 出来る 此れ 成る 型 動画 頂く ゲーム 今回 サイト</li> <li>- ユーザー 中 募集 ニュース 記事 コメント 評価 機能 レビュー 投稿 メッセージ 通知 各種 機能_利用_出来る 新規 成る 新しい わあ ザップ 攻略</li> <li>- 日 選手 位 監督 為る 年 チーム 日本 巨人 大 投手 ロッテ 試合 野球 スポーツ 回 纏め 戦 成る 今季</li> </ul>

Figure B.8: The topic keys from the Webhose's Free Datasets with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

KO-KAIST

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- 명 하다 한는 교수대 개은 중 연구 이다 받다 돼다 학교도 전 교육 대학 예서는 많다</li> <li>- 하다 적 은는 없다 수 이다 되다 아니다 않다 도 말 문제 국민 대통령 다고 노 인 하고</li> <li>- 하다 은 한 되다 대통령 의원 장관 대해 한나라당 위 열리다 적 인 정부 국회 이다는 위원회 청와대 우리당</li> <li>- 하다는 은 씨 않다도 이다 말 사람 고 보다 없다 좋다 오다 안 정도 못 가다 다시 되다</li> <li>- 하다는 북한는 한국 미국 일본 회담 중국 되다 하고 문제 적 차 측 말 인 이라고 돼다</li> <li>- 하다는 한 이다 미국 인 되다 만들다 예는 개 도는 수 인터넷 돼다 기술 세계 업체 크다 제품</li> <li>- 하다는 되다는 정부 도 수 방안 사업 한 개발 계획 따르다 이다 아파트 지역 위해 경우 형 포함 검토</li> <li>- 한 인 는 은 다 영화 도 되다 이다 적 작품 책 사랑 사진 받다 하다 한국 감독 서울작가</li> <li>- 하다는 한 되다 전 고 말는 검찰 씨 사건 수 조사 대한 도청 돼다 수사 밝히다 받다 김</li> <li>- 은 하다는 원 이다는 기업 투자 올해 되다 보다 도 세금 따르다 한국 다 소득 시장 달라 수 증가</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- 은 기업 도 이다 투자 한국 는 성 보다 국내 미국 최근 많다 만 시장 경제 다 경영 하다 새롭다</li> <li>- 명 돼다 인 대다 대 따르다 가다 은 전 출신 되다 이다 중 주 개 한국 조 많다 경우 나타나다</li> <li>- 하다는 대통령 도 정부 노 이다 정책 국민 은 아니다 한나라당 <b>열리다_우리당</b> 청와대 정권 노조 의원 대해 는 국회 문제</li> <li>- 원 는 하다는 올해 이다 정부 은 지난해 세금 아파트 따르다 부담 세 내년 가구 부동산 예산 주택 소득 거래</li> <li>- 은 하다는 이다 도 씨 다 <b>씨</b> 는 보다 영화 오다 인 아니다 사랑 그렇다 감독 가족 않다 들다 가다</li> <li>- 하다는 고 시 은 되다 도 이나 학교 논술 <b>고_말_하다</b> 하고 학생 대학 받다 교육 행사 문제 인 라며</li> <li>- 하다는 교수는 이다 <b>말_하다</b> 한 <b>교수</b> 는도 황 좋다 연구 서울대 팀 강 되다 운동 김 계 때문</li> <li>- 도청 한 하다는 씨 검찰 전 대한 수사 국정원 테이프 돼다 대해 당시 불법 조사 사건 알려지다 공 자료</li> <li>- 는 하다는 처럼 되다 은 서울 사진 속 전 집 예는 보다 작품 만들다 책 작가 만 한 미술관 서</li> <li>- 북한 회담 미국 는 중국 한국 밭다 장 예는 돼다 남북 인 날 문제 차 핵 갖다 중국 대사</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- 원 투자 올해 따르다 이다 은 도 세금 아파트 는 부동산 부담 지난해 늘어나다 소득 가구 주택 보다 내년 세</li> <li>- 씨 는 하다는 김 <b>씨</b> 는 서울 돼다 뒤 전 받다 장 인 한 들다 개 박 직원 온 예는 은행</li> <li>- 노조 개 돼다 정부 은 권 지역 지난해 통해 <b>관계자</b> 는 방안 중 경우 대해 조 계획 밝히다 건설 경 도</li> <li>- 명 한국 기업 미국 이다 대다 최근 고 출신 경영 중국 성 사업 일본 돼다 국내 신 기술 국제 제품</li> <li>- 는 하다는 보다 은 이다 도 이나 만 좋다 않다 다 그렇다 없다 가다 많다 엔 안 살다 내 만들다</li> <li>- 교수 국회 서울대 대 의원은 논술 팀 황 학생 연구 <b>열리다_우리당</b> 대한 대학 간 은 논란 교사 개정안 안</li> <li>- 정책 노 대통령 이다 도 국민 정부 은 하다가 하고 국가 아니다 청와대 경제 반 정권 이라는 한나라당 이렇다</li> <li>- 북한 회담 도청 검찰 국정원 테이프 문제 대한 장관 내용 수사 남북 한 전 대해 인 불법 은 자료 차</li> <li>- 돼다 도 는 예는 만들다 일본 참배 더 보내다 사진 총리 되다 준 휴대전화 주민 인터넷 라며 급 현지 테러</li> <li>- 점 은 영화 다 처럼 사진 사랑 인 작품 예서는 전 미술관 감독 도 오다 서울 나오다 작가 맞다 가다</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- 하다는 은 도 대통령 국민 노 이다 아니다 는 정책 청와대 다 않다 정부 하고 한나라당 문제 국가 <b>대통령_은</b> 대표</li> <li>- 북한는 회담 은 미국 날 문제 일본 밭다 하다가 남북 인 대사 차 하고 돼다 중국 대표 한국</li> <li>- 하다는 교수는 대 이다 <b>교수_는</b> 한 <b>말_하다</b> 서울대 은 황 팀 연구 인터넷 강 최근 만 운동 권 가다</li> <li>- 원 은 는 도 올해 세금 아파트 따르다 하다는 정부 내년 세 부담 부동산 가구 소득 주택 형 돼다 이다</li> <li>- 은 도청 한 는 전 하다는 검찰 씨 대한 돼다 대해 조사 <b>관계자_는</b> 수사 국정원 테이프 밝히다 사장 사건 당시</li> <li>- 하다는 이다 은 보다 <b>씨_는</b> 다 도 오다 영화 내 씨 가다 사랑 여성 그렇다 처럼 안 <b>고_말_하다</b> 사람</li> <li>- 인 는 도 사진 전 되다 은 씨 작품 점 한 서울 작가 예는 보다 예서는 속 오후 미술관 다</li> <li>- 은 기업 한국 이다 는 도 투자 지난해 인 되다 최근 국내 따르다 수 대다 경영 미국 성 일본 업체</li> <li>- 명 은 하다는 돼다 개 예서는 예는 김 이다 대다 이나 도 보이다 나타나다 서울 나오다 한 대부분 위원 앞</li> <li>- 는 정부 하다는 국회 노조 의원 방침 도 따르다 시 이다 논술 학생 장관 일부 안 위원회 되다 고 간</li> </ul>

Figure B.9: The topic keys from the KAIST Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

## TH-Prachathai

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- คดี ตัว ศาล ข้อ กฎหมาย หา สิ่ง ฟ้อง จำเลย โทษ กรณี เจ้าหน้าที พิพากษา ยื่น พิสิทธิ์ มาตรา พิจารณา สอบสวน ดำเนิน</li> <li>- งาน ศึกษา ทำ คน ยา เรียน เด็ก สุขภาพ ระบบ ปัญหา ปี โรงเรื่อง ประกัน ชาย หลีก บริการ ป่วย โรค พัฒนา</li> <li>- คน ชุมชนم เจ้าหน้าที รก ทหาร ตำรวจ เวลา พื้นที่ เหตุการณ์ ตัว บ้าน น. ชีวิต บริเวณ ชาว ยิง หน้า แดง ระเบิด เลือ</li> <li>- ประชาชน เมือง สิทธิ รัฐบาล สังคม คน อำนาจ ไทย ทำ ประชาธิปไตย ปัญหา รัฐ ประเทศ ข้อ เรียกร้อง สร้าง กฎหมาย เสรีภาพ รัฐประหาร รุนแรง</li> <li>- พม่า ประเทศ รัฐบาล ไทย ปี ทหาร กอง คน เมือง ชาว จีน ตัว ทำ รายงาน ประท้วง ทัพ ระบุ สหรัฐฯ โลก ชาติ</li> <li>- สื่อ ข่าว ข้อมูล งาน ทำ เรื่อง บริการ เว็บไซต์ ข้อ รายการ เสนอ กรณี รายงาน กิจการ กฎหมาย ระบุ มวล อนุญาต ชน วิทย์</li> <li>- บ้าน พื้นที่ น้ำ โครงการ ชุมชน ทำ ปัญหา ที่ดิน ปา งาน ไฟฟ้า กระทบ สร้าง โรง แวดล้อม ดำเนิน เขต พัฒนา ประชาชน ก่อสร้าง</li> <li>- คน ทำ ผม เรื่อง ตัว ดี ระบุ ปี ดู งาน งาม ท่าน เหมือน ไทย ชีวิต รู้สึก ตอน ภาพ บ้าน เวลา</li> <li>- งาน เงิน ค่า ทำ แรงงาน บาท ปี คน ล้าน จ้าง ทุน ประเทศ ร้อย บริษัท ไทย ราคา เศรษฐกิจ เดือน ลด จ่าย</li> <li>- รัฐมนตรี คณะ นายกรัฐมนตรี ร่าง เลือกตั้ง พรรค เรื่อง กฎหมาย ประชุม ทำ กรรมการ ประธาน พิจารณา คน เมือง ประชาชน เสนอ หน้าที ตำแหน่ง</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- คดี ศาล จำเลย โทษ กรณี สิ่ง ฟ้อง กฎหมาย คน พิพากษา หมาย พยาน <b>ดำเนิน_คดี</b> ทำ ร้อง ทนายความ <b>ข้อ_หา</b> ทำ เจ้าหน้าที ระบุ</li> <li>- เรื่อง พิจารณา ประชุม กฎหมาย กรณี ร่าง ดำเนิน มาตรา ประกาศ รายงาน มติ พ.ร.บ. ตรวจสอบ ระบุ เสนอ ชอบ ชี้แจง ทำ ฉบับ ประธาน</li> <li>- คน ผม ทำ ตัว เรื่อง ดี ระบุ งาม ท่าน ดู สังคม เหมือน ตอน ชีวิต เวลา รู้สึก บ้าน งาน สื่อ ภาพ</li> <li>- ประเทศ ทำ ปี เรื่อง ระบุ ไทย รายงาน สหรัฐฯ ประท้วง จีน รัฐบาล คน โลก ตัว เรียกร้อง ข้อมูล เมือง เว็บไซต์ อาเซียน สื่อ</li> <li>- ศึกษา คน สังคม ปัญหา รัฐ เด็ก พื้นที่ เรื่อง ทำ สร้าง สิทธิ เรียน รุนแรง องค์กร กิจกรรม งาน พัฒนา วิชาการ <b>สิทธิ_มนุษยชน</b> ประชาชน</li> <li>- ทำ ปี ประเทศ ระบบ บริการ ทุน เงิน ยา ลด บริโภค รัฐบาล ราคาไทย ป่วย งบ ประชาชน รัฐ นโยบาย ปัญหา โครงการ</li> <li>- แรงงาน พม่า <b>ทำ_งาน</b> ร้อย คน งาน พนักงาน จ้าง ปี บริษัท <b>คน_งาน</b> รัฐบาล ลูกจ้าง เมือง กอง ทำ รายงาน <b>ข้าม_ชาติ</b> <b>สหภาพ_แรงงาน</b> ระบุ</li> <li>- ชุมชนم เจ้าหน้าที ทหาร ชีวิต เหตุการณ์ คน พื้นที่ <b>เวลา_น.</b> ตำรวจ รายงาน เดินทาง ทำ รก บ้าน ยิง ก่อ บริเวณ ระเบิด บาดเจ็บ <b>เจ้าหน้าที_ตำรวจ</b></li> <li>- ประชาชน เมือง รัฐบาล เลือกตั้ง ประชาธิปไตย พรรค ทำ คน อำนาจ รัฐธรรมนูญ ชุมชนم ประเทศ ปัญหา เรียกร้อง กฎหมาย ไทย รัฐประหาร สร้าง ชัดแย้ง สังคม</li> <li>- บ้าน พื้นที่ ชุมชน โครงการ ทำ ปัญหา กระทบ น้ำ แวดล้อม ที่ดิน ประชาชน สร้าง ดำเนิน ประชุม บริษัท เครือข่าย คน จังหวัด ปา คัดค้าน</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- กฎหมาย เรื่อง เลือกตั้ง ประชุม พิจารณา รัฐธรรมนูญ พรรค ประชาชน <b>นายกรัฐมนตรี</b> มาตรา ร่าง กรณี คสช. มติ กกต. เมือง ทำ ประกาศ ประธาน เสนอ</li> <li>- ประเทศ พม่า รายงาน ระบุ ปี ประท้วง รัฐบาล เมือง สหรัฐฯ จีน ทำ เรื่อง ไทย กอง ทหาร อาเซียน เรียกร้อง ชาว คน กัมพูชา</li> <li>- สื่อ ข้อมูล ข่าว เรื่อง เว็บไซต์ ศึกษา กรณี บริการ ทำ มหาวิทยาลัย ระบุ เผยแพร่ กสทช. ข้อความ งาน สาธารณะ ภาพ เนื้อหา <b>สื่อ_มวลชน</b> รายการ</li> <li>- บ้าน พื้นที่ ชุมชน โครงการ ที่ดิน ปัญหา กระทบ น้ำ ทำ แวดล้อม ประชาชน ประชุม สร้าง เครือข่าย ดำเนิน ปา บริษัท รัฐ จังหวัด คัดค้าน</li> <li>- แรงงาน ร้อย ประเทศ เงิน ทุน บริษัท จ้าง ไทย บาท ปี พนักงาน ทำ รัฐบาล เศรษฐกิจ ลด ราคา ผลิต งาน <b>ทำ_งาน</b> เจริญ</li> <li>- ประชาชน เมือง รัฐบาล ประชาธิปไตย อำนาจ สังคม ทำ ปัญหา รัฐ สร้าง เรียกร้อง รุนแรง ประเทศ คน กฎหมาย ชัดแย้ง เลือกตั้ง ไทย สิทธิ เรื่อง</li> <li>- คน ผม ทำ เรื่อง ตัว ระบุ งาม ดี ท่าน ดู เหมือน สังคม รู้สึก ตอน ชีวิต เวลา บ้าน โลก ตอน หน้า</li> <li>- คดี ศาล จำเลย สิ่ง โทษ กรณี กฎหมาย พิพากษา ฟ้อง หา หมาย ยื่น <b>ดำเนิน_คดี</b> พยาน ร้อง เจ้าหน้าที ทนายความ <b>ข้อ_หา</b> คน จับกุม</li> <li>- เด็ก ระบบ ยา ศึกษา สิทธิ ทำ ป่วย บริการ ปัญหา รักษา พัฒนา สุขภาพ เรื่อง แพทย์ ปี ประชาชน ประเทศ ประเทศไทย หญิง เรียน</li> <li>- ชุมชนم เจ้าหน้าที เหตุการณ์ ทหาร พื้นที่ <b>ชีวิต</b> <b>เวลา_น.</b> คน ตำรวจ รายงาน รก เดินทาง บริเวณ ยิง ก่อ ระเบิด บ้าน ทำ หน้า บาดเจ็บ</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- บริการ ยา ระบบ งาน ทำ สุขภาพ หลีก ข้อมูล เรื่อง ป่วย ประกัน โรค รักษา ประชาชน สิทธิ ติด ปัญหา หน่วย ปี พัฒนา</li> <li>- ทหาร พม่า พื้นที่ กอง ชาย แดน คน จังหวัด ไทย บ้าน ทัพ งาน เจ้าหน้าที รัฐบาล ชาว ทำ ปี ตัว กัย รายงาน</li> <li>- คน ชุมชนม รก เจ้าหน้าที ตำรวจ เวลา น. เลือ แดง บริเวณ ชีวิต ตัว หน้า ข่าว เหตุการณ์ ทหาร ระเบิด ยิง รายงาน ทำ</li> <li>- งาน เงิน ค่า ทำ แรงงาน บาท คน ปี ล้าน ทุน จ้าง ร้อย บริษัท ประเทศ เดือน เศรษฐกิจ ราคา ไทย ลด จ่าย</li> <li>- ประชาชน เมือง สิทธิ รัฐบาล อำนาจ คน สังคม ทำ ไทย ประชาธิปไตย ปัญหา รัฐ สื่อ ข้อ ประเทศ เรียกร้อง กฎหมาย สร้าง ชน เสรีภาพ</li> <li>- คดี ตัว ศาล ข้อ กฎหมาย สิ่ง หา ฟ้อง จำเลย โทษ เจ้าหน้าที กรณี พิพากษา ดำเนิน ยื่น สิทธิ สอบสวน มาตรา พิจารณา ปี</li> <li>- คน ทำ เรื่อง ผม ตัว ดี ศึกษา เรียน งาน ระบุ เด็ก ปี ดู สังคม งาม ชีวิต ไทย หญิง ท่าน เหมือน</li> <li>- ประเทศ ปี รัฐบาล ไทย ทำ เรื่อง โลก ชาว จีน ระบุ ตัว คน ประท้วง สหรัฐฯ รายงาน เมือง สื่อ ประเทศไทย อาเซียน เจริญ</li> <li>- บ้าน พื้นที่ น้ำ โครงการ ชุมชน ทำ ปัญหา ที่ดิน งาน ปา ไฟฟ้า กระทบ โรง สร้าง แวดล้อม พัฒนา ดำเนิน เขต ประชาชน ก่อสร้าง</li> <li>- คณะ รัฐมนตรี นายกรัฐมนตรี ร่าง รัฐธรรมนูญ เลือกตั้ง เรื่อง พรรค ประชุม ทำ กฎหมาย กรรมการ ประธาน พิจารณา เสนอ คน งาน ข้อ หน้าที ประชาชน</li> </ul>

Figure B.10: The topic keys from the Prachathai with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

TH-Wongnai

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- คน สั่ง รอ ร้าน โต๊ะ พนักงาน ทำ นั่ง เดิน ตอน ทาน ผม ดู เวลา เหมือน ลูกค้า บริการ คิว ตัว เค้</li> <li>- เนื้อ ดี ทาน อร่อย น้ำ สลัด เลือก ผัก ซอส บาท ราคา งาน หมู เมนู อาหาร สั่ง ญี่ปุ่น สด ข้าว ซอบ</li> <li>- ร้าน นั่ง กาแฟ ดี บรรยากาศ ดีม เครื่อง เย็น ตกแต่ง อาหาร รัก เหมาะ เด็ก สวย เลือก สั่ง ซอบ เล่น หวาน รสชาติ</li> <li>- อาหาร ร้าน ดี ทาน ราคา รสชาติ บริการ เมนู คน ผม สำหรับ อร่อย ทำ สาขา บรรยากาศ ดู ไทย เลือก งาน แพง</li> <li>- ทาน ทาน อร่อย ดี รส รสชาติ ร้าน บาท น้ำ ขนม หอม ตัว เมนู นม ชื่น ซอบ เลือก เด็ก ขนมปัง ลอง</li> <li>- ร้าน รก จอด ถนน ขาย ซอย ราคา หน้า บาท นั่ง รีม หา ติด เดิน แถว ขับ มือ ผัง โต๊ะ ซ้าย</li> <li>- กิน อร่อย ร้าน คน ลอง ซอบ สั่ง ดี ดู เค้ เหมือน ชิม เพื่อน แวะ ตอน รู้ ก้อ ราคา รสชาติ เน้นน้ำ</li> <li>- หมู ร้าน น้ำ อร่อย ก้วยเดียว เส้น ข้าว ทาน ซุป สั่ง บาท เนื้อ ใส ดี เปิด ซาม ไข่ ชื่น นุ่ม รสชาติ</li> <li>- ผัด อาหาร ปลา กุ้ง อร่อย ดี น้ำ สด ทอด งาน รสชาติ ปู ตัว สั่ง ข้าว เมนู ทาน ทะเล ร้าน ไข่</li> <li>- ไข่ อร่อย ร้าน น้ำ หมู ทาน ข้าว ดี ทอด จิ้ม รสชาติ ย่าง อาหาร เมนู ผัด สั่ง งาน รส ซอบ แยก</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- อร่อย ผัด กุ้ง ทอด ดี สด สั่ง งาน ปลา อาหาร รสชาติ เมนู ปู ผัด ตัว ทาน ยา ร้าน แถว กรอบ</li> <li>- คน สั่ง รอ ผม ร้าน โต๊ะ พนักงาน ทำ ดี ดู กิน เดิน ตอน ทาน ลูกค้า นั่ง อาหาร บริการ คิว ตัว</li> <li>- ทาน อร่อย เมนู งาน สลัด ดี ซอส รสชาติ กรอบ ไข่ ผัก สั่ง ร้าน ทอด น้ำ หอม ซอบ แป้ง นุ่ม หมู</li> <li>- หวาน ทาน อร่อย ดี รส บาท ร้าน เมนู ตัว รสชาติ หอม ชื่น เลือก ขนม ลอง นุ่ม ขนมปัง ซอบ ใส นม</li> <li>- ร้าน <b>จอด_รก</b> บาท ซอย จอด ผม หา คน ขาย ราคา ทาน แถว เดิน <b>หน้า_ร้าน</b> ป้าย โต๊ะ ติด เจอ รก บริการ</li> <li>- ร้าน นั่ง ดี กาแฟ บรรยากาศ เด็ก เย็น <b>เครื่อง_ดี</b>ม สั่ง รัก ตกแต่ง อร่อย ลอง รสชาติ ซอบ ขนม หวาน แก้ว ร้อน เมนู</li> <li>- ดี ร้าน ทาน อร่อย เนื้อ เลือก ราคา สาขา ข้าว สั่ง สด กิน ย่าง หมู บาท เมนู ญี่ปุ่น <b>น้ำ_จิ้ม</b> ซอบ ซุด</li> <li>- กิน อร่อย ร้าน ลอง คน ซอบ สั่ง เค้ ดี ดู แวะ ชิม ก้อ เพื่อน ทาน เน้นน้ำ รู้ แถว ตอน เหมือน</li> <li>- ร้าน อาหาร ดี บรรยากาศ บริการ นั่ง ราคา รสชาติ <b>ร้าน_อาหาร</b> ทาน เมนู หลากหลาย เลือก ตกแต่ง สด ดู สวย พนักงาน ห้อง เบียร์</li> <li>- หมู ร้าน อร่อย ข้าว ก้วยเดียว ไข่ เส้น ทาน สั่ง น้ำ ดี เปิด ซาม <b>น้ำ_ซุ</b>ป นุ่ม ชื่น บาท รสชาติ ใส เนื้อ</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- ทาน ทาน อร่อย รสชาติ รส ตัว ดี ร้าน บาท ชื่น หอม เมนู นุ่ม ซอบ เลือก ขนม แป้ง ลอง ใส</li> <li>- ร้าน กาแฟ ดี นั่ง เด็ก หวาน <b>เครื่อง_ดี</b>ม เย็น รสชาติ สั่ง ซา อร่อย แก้ว ลอง เมนู นม ขนม รัก บรรยากาศ ซอบ</li> <li>- ดี น้ำ อร่อย ทาน เนื้อ ร้าน เลือก ราคา ข้าว สด ซุป จิ้ม สั่ง บาท เมนู ซอบ กิน ซุด หน้า ญี่ปุ่น</li> <li>- หมู ร้าน น้ำ อร่อย ข้าว ก้วยเดียว เส้น ทาน สั่ง ดี ซุป ไข่ ใส รสชาติ เนื้อ เปิด ซาม นุ่ม บาท เครื่อง</li> <li>- ร้าน อาหาร บรรยากาศ ดี นั่ง บริการ รสชาติ ราคา <b>ร้าน_อาหาร</b> ตกแต่ง เลือก อร่อย เมนู สวย ทาน โชน เหมาะ แอร์ หลากหลาย ห้อง</li> <li>- คน ร้าน สั่ง ผม รอ ทาน พนักงาน โต๊ะ ดี อาหาร บริการ ทำ นั่ง ลูกค้า เดิน ตอน สาขา ดู เหมือน เวลา</li> <li>- ร้าน <b>จอด_รก</b> บาท ซอย หน้า หา จอด แถว ราคา ถนน ขาย ติด บ้าน ผม เดิน คน แวะ เจอ ป้าย โต๊ะ</li> <li>- กิน อร่อย ร้าน ลอง สั่ง คน ดู เค้ ซอบ ดี แวะ เพื่อน ชิม ก้อ รู้ ตอน ราคา เหมือน หา เน้นน้ำ</li> <li>- อร่อย กุ้ง ผัด ทอด สด ปลา ดี งาน รสชาติ อาหาร สั่ง ปู เมนู แถว ทาน ไข่ น้ำ ยา ตัว ร้าน</li> <li>- อร่อย ไข่ ทาน ดี ทอด หมู งาน เมนู รสชาติ สลัด ร้าน ผัก ย่าง กรอบ ซอส อาหาร สั่ง น้ำ เนื้อ นุ่ม</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- ทาน หวาน ดี อร่อย รสชาติ รส ชื่น หอม นุ่ม ตัว แป้ง กรอบ เนื้อ ซอส ลอง ทำ บาท ใส ขนม กลิ่น</li> <li>- ร้าน อาหาร ดี ทาน ราคา บริการ รสชาติ อร่อย ผม เมนู สาขา แพง พนักงาน คน ดู งาน สำหรับ ทำ บรรยากาศ เลือก</li> <li>- กิน ร้าน อร่อย ลอง ซอบ คน สั่ง ดี ดู ผม ชิม เหมือน รสชาติ เค้ เพื่อน แวะ ตอน ก้อ รู้ ราคา</li> <li>- อาหาร อร่อย ผัด ทอด ปลา กุ้ง น้ำ งาน รสชาติ สด ดี ร้าน สั่ง ทาน เมนู ข้าว ปู ผัด แถว ตัว</li> <li>- คน สั่ง รอ โต๊ะ ทำ ร้าน พนักงาน ตอน เดิน นั่ง เวลา ดู กิน ทาน ลูกค้า คิว เหมือน เค้ ดี หน้า</li> <li>- ร้าน อาหาร นั่ง บรรยากาศ ดี ดีม เครื่อง ตกแต่ง ห้อง เหมาะ สำหรับ สวย เล่น ไทย เย็น แอร์ กาแฟ รัก บริการ โรง</li> <li>- ร้าน หวาน กาแฟ น้ำ ดี บาท รสชาติ เครื่อง เมนู ดีม นม ราคา สั่ง อร่อย เด็ก เลือก ซา เย็น ขนม ซอบ</li> <li>- ร้าน รก จอด ถนน ขาย หน้า ซอย หา แถว บาท ติด เดิน ราคา รีม คน แวะ ขับ โต๊ะ บ้าน มือ</li> <li>- หมู น้ำ ร้าน อร่อย ข้าว ก้วยเดียว เส้น ไข่ ทาน ซุป สั่ง ดี บาท เนื้อ ใส รสชาติ ไข่ เปิด ซาม นุ่ม</li> <li>- เนื้อ น้ำ อร่อย ดี สลัด หมู เลือก ผัก ทาน บาท สด จิ้ม ไข่ ข้าว สั่ง ย่าง ญี่ปุ่น ราคา งาน ซุป</li> </ul>

Figure B.11: The topic keys from the Wongnai with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

## TH-BEST

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- คน ทำ เรื่อง นายกรัฐมนตรี ข่าว พรรค รัฐบาล เมือง คณะ งาน ประชาชน ประชุม สื่อ เลือกตั้ง ตัว คดี ข้อ เวลา ศาล</li> <li>- น้ำ ทำ พื้นที่ ตัว บริเวณ โครด ทะเล ท่วม ดิน อากาศ เชื้อ สาร ระดับ สัตว์ ฝน ไร่ ไร่ ตก นก ป่วย</li> <li>- เงิน ไทย ประเทศ ปี งาน เรื่อง ทำ เรียน บาท คน ทูต เด็ก ราคา หนังสือ พิมพ์ ล้าน สินค้า พ.ศ. ค่า โรงแ</li> <li>- เครื่อง ทำ ไทย ผ้า รูป สี เรือ กวายเป็น งาน สร้าง ช่าง เล่น วัด พระองค์ เพลง ละคร ไม่ เสด็จ คน</li> <li>- คน รถ บ้าน เจ้าหน้าที่ ตำรวจ ปี ตัว เวลา พื้นที่ ไร่ ทำ อายุ ระเบิด หา แจ่ง น. ชาย ชีวิต ตรวจสอบ ทราบ</li> <li>- หญิง คน หน้า ตัว เสียง สาว หลอน ชาย ดู ตา ทำ เหมือน มือ เดิน สี หนูม ห้อง นิ่ง สอง ดวง</li> <li>- อาหาร ทำ ไม้ ต้น ปลูก พืช พันธุ์ ชนิด ใบ ตัว ดิน น้ำ สัตว์ ดี เสียง ดอก ข้าว สี ยา ขนาด</li> <li>- คน ผม ทำ แม่ รู้ ดี เรื่อง บ้าน ลูก ตัว พ่อ พี่ หน้า เหมือน งาน งาม สาว ชาย ดู เรียน</li> <li>- ไทย คน เมือง ภาษา ตัว ปี ทำ เรื่อง ศาสนา สังคม โลก ชีวิต ชน อำนาจ วัฒนธรรม อินเดีย ประเทศ ปกครอง ตะวัน รุนแรง</li> <li>- ทำ สังคม ระบบ กฎหมาย ปัญหา รัฐ งาน พัฒนา ชุมชน คน สิทธิ เรื่อง ประเทศ สร้าง ศึกษา ประชาชน อำนาจ ประโยชน์ เศรษฐกิจ ไร่</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- ทำ น้ำตาล อาหาร ปลูก พืช ดิน ดี ต้น พันธุ์ ชนิด ข้าว เสียง ใบ ผลผลิต วิธี หญ้า กิ่ง ขนาด เมล็ด หมู</li> <li>- ประเทศไทย ปี ทำ ประชาชน เงิน คน ร้อย ไร่ นก ตัว เรื่อง ป่วย ราคา เดินทาง ทูต บริษัท บริการ ชื่อ</li> <li>- คน แม่ ตัว พ่อ ทำ ผม บ้าน ลูก เด็ก ท่าน หน้า รัก รู้ เรื่อง เหมือน ดี ชีวิต เสียง ไร่ ไร่ ดู</li> <li>- ทำ สังคม คน รัฐ เมือง ประเทศ เรื่อง ปัญหา กฎหมาย ระบบ ชุมชน สร้าง อำนาจ ประชาชน สิทธิ ศึกษา ประโยชน์ รู้ แนว พัฒนา</li> <li>- เรื่อง ไทย พิมพ์ ประเทศ เล่ม ศึกษา ภาษา หนังสือ อินเดีย พระองค์ ประเทศไทย เด็ก งาน สร้าง แผนที่ โครงการ สหกรณ์ ทำ จารึก</li> <li>- น้ำ ทำ บริเวณ ตัว อากาศ พื้นที่ น้ำท่วม ร่างกาย สาร ผึ้ง ชนิด ระดับ ดิน อาศัย ไวรัส ทะเล เซลล์ ลักษณะ หายใจ ชีวิต</li> <li>- เจ้าหน้าที่ บ้าน ตำรวจ คน พื้นที่ อายุ ปี คน ไร่ แจ่ง ตรวจสอบ ทำ ชีวิต ก่อ เวลา น. ระเบิด ทราบ รถ รายงาน เดินทาง เหตุการณ์ หา</li> <li>- คน เรื่อง ทำ เมือง ประชาชน รัฐบาล ประชุม กกด. เลือกตั้ง งาม พรรค กรณี นายก ศาล คดี พิจารณา กฎหมาย ปัญหา ตัว พ. ด.ท.ทักษิณ</li> <li>- ผม หลอน คน ทำ หน้า ไร่ งาม ดี ดู เดิน หญิง สาว เสียง เหมือน เรื่อง ไหม นิ่ง ทรอก บ้าน ตอบ ตัว</li> <li>- ทำ เล่น เรือ เครื่อง ลาย ผ้า ม้า ไทย ไม้ ช่าง ดักตา ชนิด ลักษณะ ถม ตัว กลาง ลวดลาย สี เรือน รูป</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- ทำ น้ำ ดิน ปลูก ต้น ชนิด พืช พันธุ์ เครื่อง ใบ ดี ข้าว ขนาด ไม้ วิธี บริเวณ ผลผลิต อาหาร หญ้า กิ่ง</li> <li>- สังคม เมือง ทำ อำนาจ คน กฎหมาย รัฐ เรื่อง สร้าง วัฒนธรรม ชุมชน แนว ศึกษา โลก ลักษณะ ระบบ รุนแรง ประชาชน สัมพันธ์ ปัญหา</li> <li>- ไร่ ทำ อาหาร ตัว โครด น้ำ นก ร่างกาย สาร ชนิด อากาศ อากาศ ป่วย สัตว์ ผึ้ง เสียง เชื้อ ม้า ไวรัส ชีวิต</li> <li>- ไทย พิมพ์ เรื่อง หนังสือ เล่ม เรือ ผ้า เครื่อง ทำ ลาย ช่าง สร้าง เล่น รัชกาล ดักตา รูป ถม ละคร แผนที่</li> <li>- ผม แม่ คน ไร่ บ้าน พ่อ ทำ ดี เรื่อง งาม เหมือน ลูก หน้า พี่ น้ำตาล ตัว เพื่อน ยาย ดู ตอน</li> <li>- เด็ก พิมพ์ เรียน วัด ปี ครอบครัว ร้อย พระบาทสมเด็จพระเจ้าอยู่หัว คน งาน ประชาชน ดี ไร่ เรียน ครู ศึกษา เวลา งาม บ้าน แพทย์ ทำ</li> <li>- หลอน คน หน้า เสียง ตัว ดู หญิง สาว ผม ทำ เดิน เหมือน มือ นิ่ง ดี ตา ชาย หนูม รู้ ห้อง ไหม สัก</li> <li>- ประเทศ ปัญหา ทูต ทำ เงิน พัฒนา เรื่อง ระบบ โครงการ ผลผลิต ชุมชน สิทธิ ประโยชน์ ประชาชน แวดล้อม สินค้า บริการ เศรษฐกิจ ที่ดิน พื้นที่</li> <li>- พื้นที่ เจ้าหน้าที่ บ้าน ตำรวจ รถ อายุ ปี แจ่ง คน ไร่ ตรวจสอบ ทำ ระเบิด ยิง ชีวิต คน เวลา น. ตัว เดินทาง ก่อ ทราบ เนื่อง</li> <li>- เรื่อง คน ทำ รัฐบาล ประชุม เมือง ประชาชน พรรค กกด. เลือกตั้ง ศาล คดี งาม กรณี นายก รัฐมนตรี นายก พิจารณา รายงาน ตรวจสอบ ปัญหา</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- สังคม คน ทำ เมือง กฎหมาย รัฐ ไทย อำนาจ ระบบ เรื่อง ตัว สร้าง สิทธิ ชุมชน ปัญหา ประเทศ โลก วัฒนธรรม ศึกษา ชน</li> <li>- ไม้ ทำ ต้น เครื่อง ปลูก ใบ พืช ดิน ชนิด ดอก พันธุ์ สี ดี อาหาร ข้าว น้ำ ขนาด ผลผลิต เนื้อ เมล็ด</li> <li>- อาหาร ตัว น้ำ ทำ โครด สัตว์ สาร ไร่ เชื้อ เสียง อากาศ ชนิด ไร่ ร่างกาย นก ป่วย หวัด ผึ้ง เลือด งาน</li> <li>- ไทย พิมพ์ ผ้า หนังสือ ทำ ตัว ภาษา เรือ กวายเป็น เรื่อง เล่น เล่ม พ. ศ. คน พระองค์ เครื่อง สร้าง ช่าง รูป สี</li> <li>- งาน ประเทศ เงิน ทำ เรื่อง ทูต ปัญหา ปี พัฒนา ศึกษา ไทย คน สินค้า ค่า บริการ โครงการ ผลผลิต ราคา เศรษฐกิจ ระบบ</li> <li>- คน รถ เจ้าหน้าที่ ตัว ปี บ้าน ตำรวจ เวลา ไร่ ระเบิด พื้นที่ อายุ ทำ แจ่ง น. ยิง หา ชาย เดินทาง ชีวิต</li> <li>- คน หน้า หญิง หลอน สาว ทำ ตัว เสียง ชาย ดู ตา ผม เหมือน เดิน ไร่ มือ หนูม ดี ห้อง นิ่ง</li> <li>- ทำ คน เรื่อง นายก ข่าว รัฐมนตรี พรรค รัฐบาล คณะ ประชาชน งาน ประชุม เมือง สื่อ คดี เลือกตั้ง ตัว ข้อ เวลา ศาล</li> <li>- คน ผม แม่ ทำ ลูก บ้าน เรื่อง ดี ไร่ พ่อ เรียน ตัว เด็ก เหมือน น้ำตาล เพื่อน งาน หน้า ชาย ยาย</li> <li>- น้ำ พื้นที่ บริเวณ ทำ บ้าน ป่า ท่วม ดิน ฝน ตก เขต ระดับ เหนือ ทะเล อากาศ ไหล หมู ตะวัน ชุมชน แหล่ง</li> </ul>

Figure B.12: The topic keys from the BEST Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022)

TH-TNC

<p><b>word</b></p> <ul style="list-style-type: none"> <li>- ผม งาน ทำ คน เรื่อง เงิน ดี ปี ขาย ตัว ชื่อ รัก เดือน ดวง บ้าน ตลาด โชค เดินทาง ตอน รถ</li> <li>- กิน คน ทำ ทาน อาหาร ร้าน ใส น้ำ ข้าว เครื่อง ตัว ขาย บ้าน ขนม ชื่อ ดี เนื้อ หา เวลา ชื่อ</li> <li>- คน ทำ หน้า ตัว ดี รู้ เดิน เหมือน ตา ดู เสียง พี่ ละ ตอน ผม ชาย เพื่อน นิ่ง เรื่อง หญิง</li> <li>- สัมภาษณ์ เรียน ทำ ศึกษา งาน คน ปัญหา ถาม ตัว ดี เรื่อง เด็ก เวลา บุคคล แนว ต้องการ โรง รู้ เกี่ยว วิธี</li> <li>- ไม้ น้ำ ภาพ ทำ ปลา บ้าน เรือ เวลา ปี ดู สี สร้าง ขนาด นก ตัว เหมือน ทะเล เมือง ชนิด สอง</li> <li>- โทษ มาตรา ข้อ คณะ กฎหมาย กรรมการ เงิน งาน กรณี รัฐมนตรี ใด ราชการ ประกาศ พระราชบัญญัติ ปี บังคับ พ.ศ. พิจารณา ทะเบียน อำนาจ</li> <li>- ลูก แม่ คน อาการ โรค เด็ก ทำ ยา ตัว หมอ บ้าน พ่อ ดิฉัน ไข้ แพทย์ กิน รักษา อาหาร ชาย ร่างกาย</li> <li>- ไทย คน ทำ ม.ร.ว.ดิศกุลธิ์ เมือง สังคม อำนาจ ปกครอง ดี ประเทศ พระมหากษัตริย์ชาติ ประชาชน รัฐบาล ม.ร.ว.ดิศกุลธิ์ เรื่อง สร้าง ปัญหา ทหาร ตัว</li> <li>- เรื่อง ภาษา ชีวิต คน ตัว รู้ เสียง โลก บท อ่าน ไทย มนุษย์ จิต แปล ท่าน หมายถึง ทำ ละคร ดี ความหมาย</li> <li>- น้ำ ท่อ ระบบ ทำ เครื่อง ขนาด ร้อน เพลิง สูบ สำหรับ ผลิต อาคาร งาน หน่วย ต้น วาจก รูป ระบาย อัตรา สร้าง</li> </ul>	<p><b>t-statistics</b></p> <ul style="list-style-type: none"> <li>- งาน เรื่อง ทำ คน เงิน ดวง รัก ดี ขาย ตัว เดินทาง ปี หมอ <b>ทำ_งาน</b> ลูก ชื่อ การงาน ประเทศ อยุ่ เรียน</li> <li>- บ้าน ภาพ ปลา ทำ นก เรือ น้ำ ชนิด ไม้ เมือง ตัว หาง ขนาด เหมือน ติด สร้าง เครื่องมือ ปาก เวลา สำหรับ</li> <li>- เรื่อง ภาษา อ่าน กรรม สร้าง แปล ความหมาย หมายถึง วาจก หน่วย ประโยค ชื่อ บท ศัพท์ ลักษณะ ทำ รูป ท่าน คน</li> <li>- คน ทำ ชีวิต เด็ก ตัว ลูก ดี รู้ รู้สึก แม่ เรื่อง อาการ มนุษย์ ดิฉัน จิต แพทย์ เวลา สุข โรค ใด</li> <li>- โทษ มาตรา ข้อ กรณี ใด ทะเบียน กฎหมาย ประกาศ พระราช บัญญัติ คำ พิจารณา ราชการ บังคับ หมายถึง เวลา ศาล เครื่องหมาย บุคคล</li> <li>- สัมภาษณ์ ทำ เรียน ท่อ ถาม ระบบ ต้องการ น้ำ วิธี ปริญญา งาน ปัญหา ศึกษา ข้อมูล บริการ ทดสอบ สำหรับ ขนาด <b>แนว_แนว</b> ลักษณะ</li> <li>- ผม คน ทำ งาน เวลา ตัว ปี เรื่อง ตอน <b>ทำ_งาน</b> เหมือน ดี ดู บ้าน ท่าน รู้ โลก ชื่อ บริษัท ชา</li> <li>- ทำ คน หน้า ตัว ดี เดิน ละ รู้ ดู เสียง ตอน เหมือน นิ่ง ถาม เรื่อง ตา พี่ หรอก เพื่อน</li> <li>- ไทย ทำ ม.ร.ว.ดิศกุลธิ์ คน เมือง <b>คน_ไทย</b> ปกครอง สังคม ประชาชน อำนาจ ม.ร.ว.ดิศกุลธิ์ ดี ประเทศ พระมหากษัตริย์ ศึกษา เมืองไทย สร้าง รัฐบาล เรื่อง ชาติ</li> <li>- กิน ทาน ทำ คน ใส ร้าน อาหาร ตัว ขาย บ้าน ข้าว อร่อย น้ำ นิ่ง ชื่อ ขนม ยา หมู ดี ตัว เครื่อง</li> </ul>
<p><b>frequency</b></p> <ul style="list-style-type: none"> <li>- สัมภาษณ์ ถาม ภาพ ปลา เรือ ตอบ ทำ บ้าน เครื่องมือ ชนิด ต้องการ ทะเล ไม้ อวน เวลา ติด เบ็ด บรรยากาศ ลักษณะ ตัว</li> <li>- คน ทำ หน้า เดิน ตัว ละ รู้ ดี ดู เสียง ตา ตอน เหมือน นิ่ง หา ถาม เพื่อน แม่ หรอก มือ</li> <li>- มาตรา โทษ ข้อ กรณี กฎหมาย ใด ทะเบียน ประกาศ พระราช บัญญัติ พิจารณา ราชการ ศาล คำ สั่ง บังคับ หมายถึง เสนอ บัญชี เวลา เงิน</li> <li>- เรียน ทำ ชีวิต ศึกษา รู้ ดี คน รู้สึก ปัญหา เรื่อง บุคคล แนว เด็ก วิธี ใด ตัว มนุษย์ ต้องการ สังคม ปริญญา</li> <li>- ไทย ม.ร.ว.ดิศกุลธิ์ ทำ เมือง <b>คน_ไทย</b> คน อำนาจ ปกครอง ม.ร.ว.ดิศกุลธิ์ สังคม ประชาชน ดี พระมหากษัตริย์ เมืองไทย รัฐบาล ประเทศ สร้าง ชาติ เรื่อง ปัญหา</li> <li>- น้ำ ท่อ ระบบ เครื่อง ทำ ขนาด อาคาร ผลิต ชนิด ต้น อากาศ ต้องการ สำหรับ อัตรา รูป ขยะมูลฝอย กัง <b>สูบ_น้ำ</b> จ่าย ลด</li> <li>- สร้าง ภาษา แปล เสียง กรรม หมายถึง นก วาจก ความหมาย หน่วย ประโยค เรื่อง อ่าน รูป ชื่อ เมือง ศัพท์ สอง คน สันสกฤต</li> <li>- ลูก ทำ คน เรื่อง เงิน หมอ งาน แม่ ดิฉัน อาการ ดวง ดี รัก โรค ตัว ขาย บ้าน ปี เดินทาง แพทย์</li> <li>- กิน ทาน ทำ คน ใส บ้าน ร้าน อาหาร ตัว ขาย ข้าว ดี ชื่อ ดู อร่อย หน้า น้ำ นิ่ง เนื้อ ขนม</li> <li>- ผม คน เรื่อง งาน ตอน ทำ ดู เวลา เหมือน ปี ตัว ดี ทาน บ้าน โลก รู้ <b>ทำ_งาน</b> อ่าน บริษัท ชื่อ</li> </ul>	<p><b>Chi-square</b></p> <ul style="list-style-type: none"> <li>- ลูก แม่ คน อาการ เด็ก โรค ทำ ยา หมอ พ่อ กิน ดิฉัน ไข้ แพทย์ ชาย บ้าน รักษา ร่างกาย หาย</li> <li>- ภาษา เรื่อง ภาพ ชีวิต ตัว กรรม เสียง บท สร้าง โลก อ่าน แปล หมายถึง ไทย ความหมาย ชื่อ ละคร มนุษย์ ท่าน สันสกฤต</li> <li>- งาน ทำ เงิน เรื่อง คน ดี ปี ขาย รัก ดวง เดือน ชื่อ ตัว โชค หุ่น ตลาด เดินทาง ประเทศ หวัง หุ่น</li> <li>- คน หน้า ทำ ตัว ดี ตา ละ เสียง เดิน รู้ เหมือน ดู พี่ เรื่อง ชาย ตอน หา หญิง นิ่ง สาว</li> <li>- สัมภาษณ์ เรียน ทำ ศึกษา งาน คน ปัญหา เรื่อง ตัว ดี ถาม บุคคล รู้ วิธี แนว ชีวิต ต้องการ เวลา รู้สึก ใด</li> <li>- โทษ มาตรา ข้อ คณะ กฎหมาย กรรมการ งาน กรณี รัฐมนตรี ใด เงิน ราชการ ประกาศ พระราชบัญญัติ พ.ศ. บังคับ พิจารณา ทะเบียน ปี อำนาจ</li> <li>- น้ำ กิน ทำ ปลา ไม้ อาหาร ทาน ใส ร้าน เครื่อง ตัว นก เรือ บ้าน ดอก เนื้อ ชนิด ขนม ปาก อร่อย</li> <li>- ผม คน ทำ งาน บ้าน ดู ปี เหมือน ตัว ตอน รถ เวลา เรื่อง ดี รู้ ท่าน สอง เดิน เพลง นิ่ง</li> <li>- ไทย คน ทำ เมือง ม.ร.ว.ดิศกุลธิ์ สังคม อำนาจ ปกครอง ดี ประเทศ ชาติ พระมหากษัตริย์ ประชาชน รัฐบาล ม.ร.ว.ดิศกุลธิ์ เรื่อง สร้าง ปัญหา ทหาร ตัว</li> <li>- น้ำ ท่อ ระบบ ทำ เครื่อง ขนาด ร้อน อาคาร สำหรับ ผลิต สูบ เพลิง งาน รูป ต้น อากาศ ระบาย อัตรา ต้องการ ชนิด</li> </ul>

Figure B.13: The topic keys from the Thai National Corpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022)



AR-ANT

<p><b>word</b></p> <p>في رئيس تونس أن حكومة على حزب يوم من نائب الذي مجلس - إلى عن سياسيي تونس وطيب وزير عام حركة في من إلى أن على ستة خلال هذا ميتة نيته الذي تونس دينار تم - تونس مليون سنهر مشروع ألف شركة يوم في أم جوهر أكد أف على مراسيل ين تصريح تونس من إف ل - مخمد تم وفق جهوي جهة مدير من منطقة يوم ولاية إلى بين في طريق و على ساعة معتقد جهة - سنارة صباح مديته ماء ترحة بعد محلي في من على أن الذي هذا عن إلى موقع هو مع أو هم عالم ذلك قد - لم صورة جديد ل - من على في أمية تم مصدر تمكّن وطيب وحده شخص حرس - إزهايب إيفاف أن ب عملية تحت أمن قد فرقة أن في على من إلى هذا الذي عن مع قانون كل وزارة قرار أضاف - أن إلى في من على قد تم الذي مستشقى حالة طغل بعد صحة - ستب هذا طيب حارت نقل عن إصاته في من أن على إلى الذي عن دولة إن إزهايب أمريكي بعد فرنسي - يوم غربي قوة ليب مع مديته عملية و في من نادي يوم فريق كرة تونس أفرقي راضي ملعب نجم - الذي مقابلة قدم ناي لايب إحد أولى ساعة</p>	<p><b>t-statistics</b></p> <p>أن في تونس من أضاف تصريح ل أكد على هذا كل اعتبر إلى - وفق هو حكومة إن بلد دولة يوم في من أو أن طغل هو صورة على هذا إن لكن إلى إن هي بعد - هاتف و قد عند الذي تونس تونس من في يوم على و أن صباح الذي شركة دولي إلى - بلاغ شوسته زبارة جوان انملق مطار عن من أمية تمكّن وحده مصدر إيفاف حرس وطيب فرقة إزهايب - تحت نيته شخص على غمومي وفق جوهر في عون تفيس أمس جوهر من ولاية في مراسيل جهة أف أم أكد جهوي معتقد وفق - منطقة إف أم شوسته مستشقى أمس على مهدي على مستوي إلى أن إن في من فرنسي إر شريطة هجوم أمريكي أمس بيان - سعودي ماضي حسب سلطة أعلن قطر ألماني مسؤول ليب في نيته من ميتة تونس أن بيعة وزارة ستة دينار إلى مشروع - مقابل على تلك ذلك ب هذا مالي حسب في يقابته أن هيئة يوم تصريح إضراب غمومي على مطلب عام - وزارة كاتب إلى إحد قرار قاضي جهوي متعلق إجراء تونس إلى من حزب حركة شاهد بوييف نداء نائب و ليب - رئيس حكومة مع سبسي رئيس جهوي رئيس مخمد في عن إن و فريق يوم نادي نجم مقابلة ملعب لايب إحد نتيجة ساجلي - على ين دار صفاقسي راضي نادي أفرقي مباراه ناني</p>
<p><b>frequency</b></p> <p>يقابته تونس كل مطلب أن و إضراب قرار عدم عام إلى غمومي - مع طالب وفق أكد أو هذا تغيير يوم تونس تونس ميتة نيته دينار مالي مشروع مقابل ستة أن تلك - مؤسسته استعمار فائده فطاع خلال حاله تقبول بين وزارة و يوم فريق نادي نجم مقابلة ملعب لايب بن إحد ساجلي - صفاقسي نتيجة جولة دار راضي ساعة سن جمعته تونس تونس أكد شاهد إن أن بوييف قساد نداء حزب نائب - سياسيي حكومة أضاف بولنيكا تغيير هيئة رئيس حكومة حول وفق مراقبه هاتف استهلاذ أضاف صحتي مائة و أو بيعة بق ذلك أوضح - أن ماء قضاء جديد إلى حقلة خاص مواطن أمين تمكّن من مصدر وحده إزهايب حرس وطيب إيفاف وفق - فرقة جوهر حزر تحت مراسيل عون غمومي بلاغ نيته جهة شريطة و صورة إن فرنسي لكن الذي أن قد أحد هذا إن على مرأة هي عبر - حث جديد بعد من فيديو مستشقى تونس وكالة حارت قد مصدر أكد إزهايب أحد أفرقيا - تحقيق إلى طيب سجن أمية أضاف اعتدا حسب جته أفاد جوهر جهوي مراسيل أكد جهة شوسته معتقد أف أم ولاية وفق - منطقة صفاقس مستسير مهدي إف أم قسر من ولاية أمس يوم تلميد ليب إن تونس أمس جزائري أعلن أن سعودي إزهايب ماضي - شريطة قطر سلطة نقل ليبيا مسؤول مصري تنظيم أورد إيطالي</p>	<p><b>Chi-square</b></p> <p>في من ستة إلى ميتة على خلال نيته أن هذا الذي دينار تم شهر - تونس مليون ألف شركة تونس مشروع رئيس في يوم حزب تونس حكومة نائب أن عن عام مجلس على - من الذي هيئة وطيب حركة جهوي سياسي انتخاب - من إلى يوم في ساعة و بعد على بين منطقة ترحة مع شركة - ساجل بحر أن شمال صباح مطار رخله من أم جوهر يوم ولاية مراسيل أف على منطقة أكد في جهة وفق - شوسته معتقد إف جهوي مديته أمية تصريح أن إلى في من على تم الذي قد مستشقى حالة حارت بعد صحة - ستب نقل عن طغل طيب و إصاته في من أن على الذي هذا عن إلى هو موقع لم أو قد ذلك هي إن إن - صورة بعد لكن من على في تم أمية أن وطيب شخص تمكّن مصدر إيفاف ب - وحده حرس قد قصيه تحت عملية إزهايب قبض و في من نادي يوم فريق تونس أفرقي راضي ين الذي ملعب - نجم مقابلة على إحد لايب كرة قدم ناني أولى في من الذي أن عن على إلى إزهايب دولة ليب يوم عشكري إن - غربي مع تونس عملية قوة تنظيم بين في أن على إلى هذا من الذي مع كل عمل عن وزارة حكومة - أضاف تونس وزير تونس قانون تم أكد</p>

Figure B.14: The topic keys from the Antcorpus with different retokenization measures. (Cheevaprawatdomrong et al., 2022)