Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 663–670

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) – Significance Sentences

Shiva Twinandilla[a], Satriyo Adhy[a], Bayu Surarso[b], Retno Kusumaningrum[a]*

[a]*Department of Informatics, Universitas Diponegoro, Semarang 50275, Indonesia*
[b]*Department of Mathematics, Universitas Diponegoro, Semarang 50275, Indonesia*

Abstract

The contents of online news documents are almost the same that will lead to the redundancy of news or called yellow journalism. Yellow journalism can make it difficult for readers to distinguish documents containing fact or opinionated information. Therefore, it is necessary to extend more research about multi-document summarization so that readers can easily understand the intent of online news documents. Latent Dirichlet Allocation (LDA) - Significance Sentences is one of the methods for summarization, which performs better than the term frequency algorithm. However, document summarization using the method is only able to summarize multiple documents as a whole without grouping by topic. Subsequently, it can give an unorganized summary result. Therefore, this research proposes a novel summarization method which combines K-Means Clustering and LDA - Significance Sentences, so it can generate document summaries based on the topic. We implemented two scenarios of the experiment. The first experimental results the best alpha value is 0.001 with the ROUGE-1 value of 0.5545 and the best summarization level is 30% with the ROUGE-1 value of 0.6118. While the second experiment results, the best obtain of ROUGE-1 value is 0.61991 for the first cluster which is consists of documents 1, 2, 3, 4, and 6 and 0.6139 for the second cluster which is consists of 5, 7, 8. Multi-document summarization using the proposed method has good performance when the K-Means method can cluster the document according to the topic correctly, which is highly dependent on the accuracy of determining the initial centroid.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the 3rd International Conference on Computer Science and Computational Intelligence 2018.

Keywords: Multi-Document Summarization; Online News; Yellow Journalism; K-Means; Latent Dirichlet Allocation; Significance Sentences; ROUGE-1

* Corresponding author. Tel.: +62-24-747-4754.
E-mail address: retno@live.undip.ac.id

1877-0509 © 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Selection and peer-review under responsibility of the 3rd International Conference on Computer Science and Computational Intelligence 2018.
10.1016/j.procs.2018.08.220

1. Introduction

In this era of globalization, science and technology will continue to grow rapidly from time to time. This causes the number of existing news documents grew, especially on the internet. Online news documents can help readers to get the latest information quickly, wherever and whenever. However, online news documents override the details and accuracy of the news because of its purpose to provide up-to-date information as much as possible. Many of the contents news documents are almost the same that will lead to redundancy of news documents or called yellow journalism. Yellow journalism can make it difficult for readers to distinguish documents containing fact or opinionated information [1]. Therefore, it is necessary to research multi-document summarization.

Research in summarization towards Indonesia document is still rarely conducted. Whereas summary can make it easier for the readers to understand the theme and concepts in the document and can shorten reading time. There are two methods of summarization, namely extractive and abstractive method [2]. The extractive method is a method for summarizing document by selecting a part of the word or sentence in the document. While the abstractive method is a method for summarizing document by creating new sentences that have the same information as the original document. In this study, we implement the extractive method for multi-document summarization.

There are some methods that have been implemented for multi-document summarization towards Indonesia document. The first research makes summarization towards twitter data based on trending issue and news important feature including word frequency, TF-IDF, sentence position and resemblance sentence to the title [3]. The second research generates summarization based on Latent Semantic Analysis (LSA) and Maximum Marginal Relevance, where the weighted variation strongly determines the accuracy of resulting summary [4]. The third research makes summarization with Latent Dirichlet Allocation (LDA) and genetic algorithm. It has been able to produce extractive summaries that includes important information from single document more quickly [5]. The fourth research proposes the used of TextTeaser Algorithm by calculating four features such as title feature, sentence length, sentence frequency, and keyword frequency [6]. The latest research generates summarization with LDA and Significance Sentence [7], which performs better than the term frequency algorithm as implemented in the previous researches. However, the method which combine LDA and Significance Sentence still has limitation, i.e. it is only able to summarize multiple documents as a whole without grouping by topic. Subsequently, it can give an unorganized summary result. Therefore, we proposed the used of K-Means combined with LDA – Significance Sentences, so it can generate document summaries based on the topic.

The rest of this paper is organized as follows. Section 2 describes related research to multi-document summarization using LDA. The detailed research procedure, as well as experimental design and results, are explained in section 3 and section 4, in sequence. Subsequently, the conclusions are drawn in Section 5.

2. Related Works

Some researches related to multi-document summarization of English document using LDA has been conducted. The first research applied an extractive method for multi-document summarization using LDA combined with mixed models for extracting topics and summarizing by taking a few sentences in the document regardless of grammatical details and document structure [8]. The second research generate multi-document summarization by selecting sentences which are obtained with a new sentence ordering method based on topic keywords using LDA [9]. The study grouped each sentence into a particular document using Similarity Histogram Clustering (SHC) and ranking the cluster using clustering importance with Sentence Information Density (SID).

The third research proposed a novel method to generate multi-document summarization with LDA and Significance Sentences [10]. The study is an improvement from their previous research using a significance topics approach to compute similarities between topics with sentences and documents [11]. The latest research using LDA in the beginning then introduces three variables include diversity of sentence distribution (α), diversity of topic distribution (β) and similarity between sentences with titles (γ) for determining sentences that have high weights as extracted sentences to form summary [10]. The advantages of LDA are (i). it is suitable for large data [12] and (ii). it performs very well in extracting topics for Indonesian text documents [13,14,15,16].

3. Research Procedure

Illustration of the research procedure can be seen in Fig 1 and the details can be seen in the following subsection.

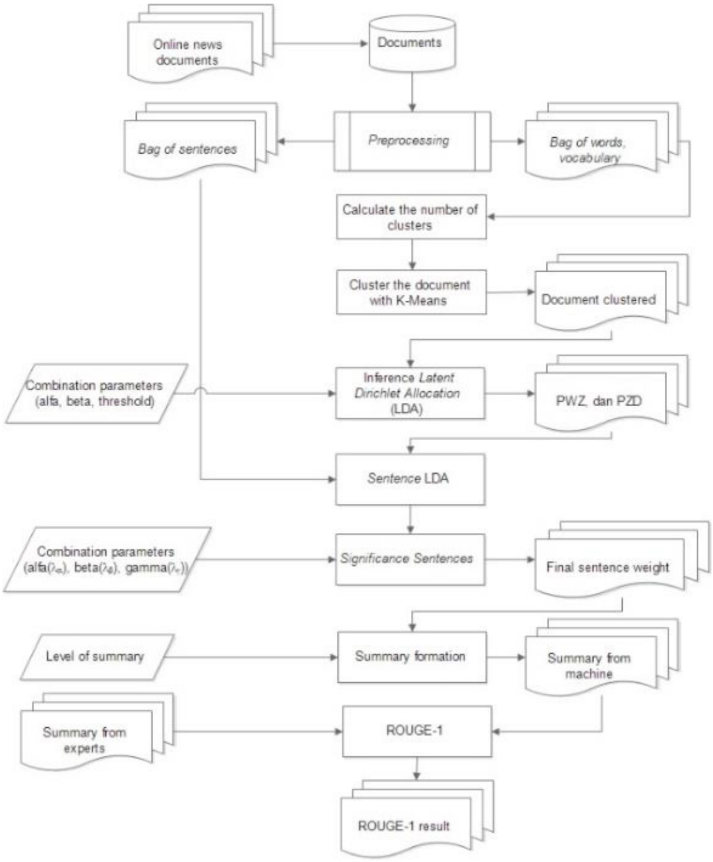

Fig. 1. Research Procedure

3.1. Preprocessing

Preprocessing is an initial stage that aims to simplify the words and sentences that exist in the document that will form a vocabulary, bag of words, and bag of sentences. Vocabulary is a collection of unique words, bag of words is a collection of words in each document, and bag of sentences is a collection of sentences in each document. This stage has four sub-processes:

- Case folding is the first process in preprocessing by converting all the letters in the document into lowercase.
- Tokenization is the second process in preprocessing, this stage has aims to cut a document into an array of words, array of sentences and eliminate punctuation. The process to form a bag of words, vocabulary and bag of sentences.
- Stopword removal is the process used to delete words that often appear but does not contain a particular meaning.
- Stemming is the process used to change the words in a document into basic word.

3.2. Calculate the Number of Clusters

Since we implement K-Means Clustering for grouping the documents into the same tropic, then we need to initially determine the number of clusters based on the following equation [17]:

$$C = \sqrt{\frac{M}{2}} \tag{1}$$

Where C is defined as the number of clusters and M is the number of the documents.

3.3. Performing K-Means clustering

Based on the selected number of clusters, the next step is to perform K-Means clustering towards the dataset, so we get some clusters of documents. Furthermore, the LDA inference process will be applied to each cluster of documents or in other words the LDA inference process is performed as much as the number of clusters. The detail algorithm of K-Means Clustering can be seen in [17].

3.4. Inference Latent Dirichlet Allocation (LDA)

LDA is the simplest topic modeling and one of the most popular probabilistic topic models [18]. LDA can apply into two processes, there are generative LDA and Inference LDA [19]. Generative LDA used to form a document from a collection of topics, whereas inference LDA used to retrieve information from a document. The use inference LDA in this study is Online Variational Bayes. Input this stage is alpha, eta, the number of topics, and a clustered bag of words. This stage will be processed before the number of change γ_{mk} divide the number of topics < 0.00001. New topics for the word selected based on the biggest value of topics. Based on the sampled topic for each word, the next step is to calculate the probability of word-topic (PWZ) based on equation 2 and topic proportions (PZD) based on equation 3.

$$\varphi_k = p(w = t|z = k) = \frac{n_{t,k}+\beta_t}{\sum_{t=1}^{V} n_{t,k}+ \beta_t} \tag{2}$$

$$\theta_d = p(z = k|d) = \frac{n_{d,k}+\alpha_k}{\sum_{k=1}^{K} n_{d,k}+ \alpha_k} \tag{3}$$

φ_k is defined as probability of word-topic or PWZ, $n_{t,k}$ is number of words t specified as topic k, β_t is hyperparameter value of alpha for probability word-topic, θ_d is topic proportions or PZD, $n_{d,k}$ is number of words in d document included in the topic k and α_k is hyperparameter value of eta for topic proportions.

3.5. Sentence LDA

Sentence LDA is LDA method approach that represents documents as topic representations and each topic is a sentence distribution that represents a sentence that has significant weighting selected on multi-document summarization [10]. Sentence LDA can be calculated with the following equation:

$$P(S_r|T_k) = \frac{\sum_{W_i \in S_r} P(W_i \mid T_k) * P(T_k \mid D_m)}{length(S_r)} \tag{4}$$

Where $P(S_r|T_k)$ is defined as probability of sentence to r on the k topic, $P(W_i \mid T_k)$ is probability of word to i on the k topic, $P(T_k \mid D_m)$ is probability of topic to k on the m document and $length(S_r)$ is number of r sentences (number of words in a sentence).

3.6. Significance Sentences

Significance Sentences is documented summarization method where each sentence has three different parameters, that are diversity of sentence distribution (α_r), diversity of topic distribution (β_r) and similarity between sentences and title (γ_r) [10].

a. Diversity of sentence distribution (α_r)
Diversity of sentence distribution using Kullback-Leibler Divergence sentence probability against topic T_k with T_K based on the following equation:

$$\alpha_r = KL(P(S_r|T_k)||P(S_r|T_K) = \sum_{t=1}^{K} P(S_r| T_k) \log_2 \frac{P(S_r|T_k)}{P(S_r|T_K)} \tag{5}$$

b. Diversity of topic distribution (β_r)
Diversity of topic distribution using Kullback-Leibler Divergence probability topic T_k document D_m dan topic T_k document D_M based on the following equation:

$$\beta_r = KL(P(T_k|D_m)||P(T_k|D_M) = \sum_{t=1}^{M} P(T_k|\, D_m) \log_2 \frac{P(T_k|D_m)}{P(T_k|D_t)} \tag{6}$$

c. Similarity between sentences and title (γ_r)
Titles of documents are an important factor in sentences selection. We evaluate the similarity between sentences and titles using KL divergence and distribution of sentences over the topic. But KL divergence is not symmetric, thus the similarity of topic distribution is used Jensen-Shannon distance [10]. The similarity between sentences and titles based on the following equation:

$$\begin{aligned}\gamma_r &= \frac{1}{2}[KL(P(T_i|T_k)||\,(P(T_i|T_s)) + KL((P(T_k|T_i)||\,(P(T_i|T_s))] \\ &= \frac{1}{2}\,[\sum_{t=1}^{K} P(W_i|\,T_k)\ \log_2 \frac{P(W_i|T_k)}{P(W_k|T_t)} + \sum_{t=1}^{K} P(W_k|\,T_i) \log_2 \frac{P(W_k|T_i)}{P(W_i|T_t)}\,]\end{aligned} \tag{7}$$

The results calculation of significance sentences obtained the value of alpha, beta, and gamma in each sentence, then done normalization to make the value between 0 to 1. The equation for normalization based on the following equation:

$$Q_r = (\alpha_r, \beta_r, \gamma_r)$$

$$Q_r = \frac{Q - Q_{min}}{Q_{max} - Q} \tag{8}$$

Where Q_r is defined as sentence parameter. Final sentence weight obtained by multiplying alpha, beta and gamma values with their parameter values [10]. Final sentence weight can calculate based on the following equation:

$$\Psi_r = \lambda_\alpha \alpha_r + \lambda_\beta \beta_r + \lambda_\gamma \gamma_r \tag{9}$$

Subsequently, Ψ_r is defined as final sentence weight, λ_α is alpha combination parameter, λ_β is beta combination parameter and λ_γ is gamma combination parameter.

3.7. Summary Formation

For each document, we sort by decreasing the value of final sentence weight, Ψ_r. Furthermore, we choose p percent of the sentences with the highest Ψ_r for each document. The p value is subsequently called as summarization level. The final step is to arrange the selected sentences in sequence.

3.8. ROUGE-1

In this study, we use ROUGE which stands for "Recall-Oriented Understudy for Gisting Evaluation" as the evaluation metric. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans [20]. In this study, summaries by human made by people with Indonesia literary background so as to maintain objectivity. Specifically, we are using ROUGE-1 for obtaining the performance of multi-document summarization.

4. Experimental and Evaluation

This research has two experiments and each experiment use different online news documents. In this study, we use two different datasets. It is because this research wants to know the performance of LDA - Significance Sentence when applied to summarize the online news document in the Indonesian language when it is compared with the proposed method which is a combination of LDA - Significance Sentence with K-means clustering for online news document in the Indonesian language that shows yellow journalism.

The first experiment uses 13 online news documents with three different topics. Aim of the first experiment is to get the best value of alpha and level summary. The relationship between the change in the value of alpha and value of ROUGE-1 can be seen in Fig. 2 and Table 1.

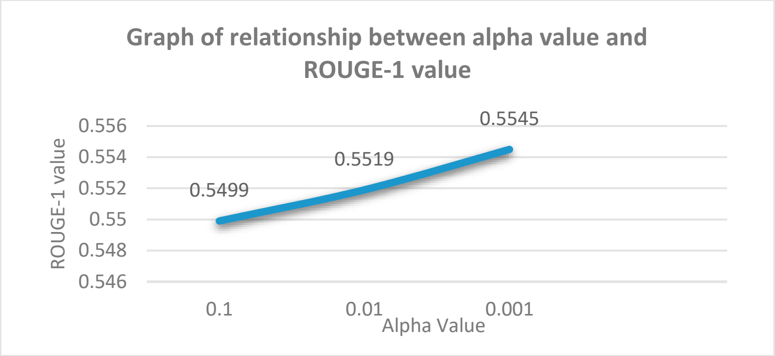

Fig. 2. The relationship between alpha value and ROUGE-1 value

Table 1. The ROUGE-1 value for each alpha value

Alpha	ROUGE-1
0.1	0.5499
0.01	0.5519
0.001	0.5545

Both Fig. 2 and Table 1 show that the best ROUGE-1 value is obtained at alpha value is 0.001, i.e. ROUGE-1 value of 0.5545. In addition, increasing the alpha value will have an effect on increasing the value of ROUGE-1. A very small alpha value will cause sparse topic distribution, as well as for the value of eta.

Furthermore, the relationship between the change of level of summarization and the ROUGE-1 value can be seen in Fig. 3 and Table 2.

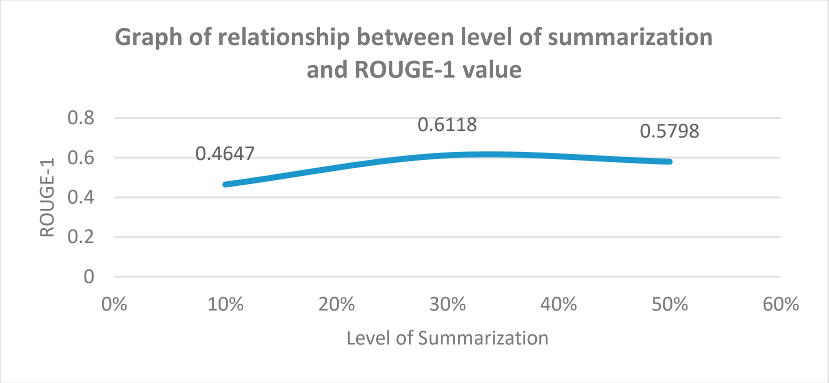

Fig. 3. The relationship between level of summarization and ROUGE-1 value

Table 2. The ROUGE-1 value for each summarization level

Level of Summarization (Percent)	ROUGE-1
10%	0.4647
30%	0.6118
50%	0.5798

Table 2 shows the value of summarization level at 10% has a ROUGE-1 value is 0.4647, at 30% has ROUGE-1 value is 0.6118 and at 50% has ROUGE-1 value is 0.5798. The highest value of ROUGE-1 is not obtained at the lowest level of summarization since the lowest level of summarization will more and more eliminate sentences that can lead to the high information lost. However, the high level of the summarization also does not guarantee yields a good

ROUGE-1 value, since it generates a large number of extracted sentences contained some unimportant information and take a long time to get information.

The second experiment uses eight online news documents with a same general topic about the death of Hari Darmawan (founder of the biggest department store in Indonesia). This research uses the alpha value of 0.1, eta value of 0.1, the threshold value of 0.001 on LDA inference process and level of summarization by 30% as resulted in the first experiment. The result of the second experiment can be seen in Table 3.

Table 3. The result of the second experiment

Process	Cluster	Member of Cluster (ID Document)	ROUGE-1
1	1	1,2,3,4,6	0.6139
1	2	5,7,8	0.6199
2	1	2	0.5833
2	2	1,3,4,5,6,7,8	0.4542

The second experiment was run twice to determine the effect of clustering process with K-Means method (for grouping the document based on its specific topic) with the obtained ROUGE-1 value. Each process produces two clusters. The first process produces cluster 1 consisting of documents 1, 2, 3, 4, 6 with ROUGE-1 value of 0.6139 and cluster 2 consisting of 5, 7, 8 with ROUGE-1 value of 0.6199, while the second process produces cluster 1 which consists of document 2 with a ROUGE-1 value of 0.5833 and cluster 2 consists of documents 1, 3, 4, 5, 6, 7, 8 with a ROUGE-1 value of 0.4542. Based on ROUGE-1 value and topic members of cluster, the first process more accurate for K-Means clustering than the second process. The first process has a pretty good result because the process of clustering document generates the appropriate document for both specific topics, i.e. the first cluster discusses the topic of the evacuation process and the funeral process of the victim Hari Darmawan, while the second cluster discusses the investigation of the causes of death Hari Darmawan.

5. Conclusion and Future Works

The proposed novel summarization method which combines K-Means Clustering and Latent Dirichlet Allocation (LDA) – Significance Sentences gives good summarization result for each specific topic when the K-Means Clustering performs better for grouping documents based on its specific topics. It is shown by the high value of ROUGE-1 for both resulted clustered, i.e. 0.61991 for the first cluster and 0.6139 for the second cluster. On contrary, it will give the bad ROUGE-1 value when the K-Means Clustering cannot correctly group the documents into each specific topic. Other parameters which are observed in this study are alpha value and level of summarization. The best alpha value is 0.001 with the ROUGE-1 value of 0.5545, while the best summarization level is 30% with the ROUGE-1 value of 0.6118. For future research, we can try other clustering methods in order to distinguish online news document based on its specific topic so that users of the system can get summarized information based on the particular topic.

Acknowledgements

The authors would like to acknowledge the research funding supported by Universitas Diponegoro under the grant of research for international scientific publication – Year 2018

References

1. Maharaj K. Technology and the Fake News Phenomenon. In Jamaica Conference Centre; 2017 Aug; Jamaica.
2. Chang YL, Chien JT. Latent Dirichlet Learning for Document Summarization. In Proceeding of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing; 2009; Taipe, Taiwan. p. 1689-1692.
3. Hayatin N, Fatichah C, Purwitasari D. Sentence Scoring based on News Features and Trending Issues for Multi-Document Summarization. Jurnal Ilmiah Teknologi Informasi. 2015 Januari; 13(1): p. 38-44.
4. Irawan S, Hermawan , Samsuryadi. Preliminary Study of Indonesia Document Summarization using Latent Semantic Analysis and Maximum Marginal Relevance. In Proceeding of Annual Research Seminar (ARS);

2016; Palembang, Indonesia. p. 235-239.

5. Silvia , Rukmana P, Aprilia RV, Suhartono D, Meiliana RW. Summarizing Text for Indonesian Language by Using Latent Dirichlet Allocation and Genetic Algorithm. In Proceeding of Electrical Engineering,Computer Science and Informatics (EECSI) 2014; 2014; Yogyakarta: EECSI. p. 148-153.
6. Gunawan D, Pasaribu A, Rahmat RF, Budiarto R. Automatic Text Summarization for Indonesian Language Using TextTeaser. IOP Conference Series: Material Science and Engineering. 2017; 190(1): p. 012048.
7. Widjanarko A, Kusumaningrum R, Surarso B. Multi Document Summarization for the Indonesian Language Based on Latent Dirichlet Allocation and Significance Sentence. In Proceeding of 2018 International Conference on Information and Communications Technology (ICOIACT); 2018; Yogyakarta. p. 520-524.
8. Arora R, Ravindran B. Latent Dirichlet Allocation Based Multi-Document Summarization. In Proceedings of The Second Workshop on Analytics for Noisy Unstructured Text Data; 2008; Singapore. p. 91-97.
9. Lukmana I, Swanjaya D, Kurniawardhani A, Arifin AZ, Purwitasari D. Multi-Document Summarization Based On Sentence Clustering Improved Using Topic Words. Jurnal Ilmiah Teknologi Informasi. 2014; 12(2): p. 1-8.
10. Na L, Ying L, Xiao-jun T, Hai-wen W, Peng X, Ming-xia L. Multi-document Summarization Algorithm based on Significance Sentences. In Proceeding of 2016 Chinese Control and Decision Conference (CCDC); 2016; Yinchuan, China: IEEE Control and Decision Conference. p. 3847-3852.
11. Na L, Xiao-jun T, Ying L, Ming-xia L, Hai-wen W, Peng X. Topic-Sensitive Multi-document Summarization Algorithm. In Proceeding of 2014 Sixth International Symposium on Parallel Architectures, Algorithms and Programming; 2014; Beijing, China: Sixth International Symposium on Parallel Architectures, Algorithms and Programming. p. 69-74.
12. Liu Z. High Performance Latent Dirichlet Allocation for Text Mining London: Brunel University School of Engineering and Design PhD Theses; 2013.
13. Prihatini PM, Suryawan IK, Mandi IN. Latent Dirichlet Allocation Method for Extracting Document's Topic. Jurnal Logic. 2017 November; 17(3): p. 154-158.
14. Kusumaningrum R, Adhy S, Wiedjayanto MIA, Suryono. Classification of Indonesian News Articles based on Latent Dirichlet Allocation. In Proceeding of 2016 International Conference on Data and Software Engineerin; 2016; Bali, Indonesia.
15. Bashri MFA, Kusumaningrum R. Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization. In Proceeding of 2017 Fifth International Conference on Information and Computational Technology (ICoICT); 2017; Melaka, Malaysia.
16. Putri IR, Kusumaningrum R. Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia. Journal of Physics: Conference Series. 2017; 801(1): p. 012073.
17. Kusumaningrum R, Farikhin. An Automatic Labeling of K-means Clusters based on Chi-Square Value. Journal of Physics. 2017; 801(1): p. 012071.
18. Blei DM. Probabilistic Topic Models. Communications of The ACM. 2012; 4: p. 77-84.
19. Kusumaningrum R, Wei H, Manurung R, Murni A. Integrated visual vocabulary in latent dirichlet allocation-based scene classification for IKONOS image. Applied Remote Sensing. 2014 January; 8(1): p. 083690.
20. Lin CY. ROUGE : a Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 Workshop - Text Summarization Branches Out; 2004; Barcelona, Spain. p. 74-81.