EasyChair Preprint

№ 1410

Create List of Stopwords and Typing Error by TF-IDF Weight Value

Woo-Seok Choi, Ki-Cheol Yoo and Sang-Hyun Choi

August 24, 2019

Create List of Stopwords and Typing Error by TF-IDF Weight Value

Woo-seok Choi[1], Ki-cheol Yoo[2], Sang-hyun Choi[2],

[1] Department of Bigdata, Chungbuk National University,
Cheongju, South Korea
[2] Department of Manangemnt Information System, Chungbuk National University,
Cheongju, South Korea
{cdt3017}@naver.com, {ryugami07, chois}@cbnu.ac.kr

Abstract. On these days, development of SNS generate huge text data. It is most important things to remove the meaningless words, stopwords and typing error to analyse text data. In English, it grew rapidly to create stopwords dictionary. However, there are few researchs in Korea for Korean language. In this research, we suggest way to firter stopwords and typing errors out by words importance with TF-IDF algorithm. First, calculate TF-IDF value from collected data. Second, decide criteria to separate to two groups by TF-IDF value and transform to n × 2 matrix. Third, calculate accumulative frequency of TF-IDF weight. In this way, new accumulative frequency is gotten without stopwords and typing error. Furthermore, this method can be used in both language : Korean and English. without creating stopwords dictionary.

Keywords: TF-IDF, Text Mining, Stopwords, Preprocessing,

1 Introduction

1.1 Purpose and Background of Research

In this research, we suggest efficient way to filter out meaningless word such as stopword, typing error, article, preposition, postposition, conjunction and just often used word by its data type in text mining.

In text mining, researchers don`t analyse normal structured data, but analyse unstructured data in many kinds of format. That`s why researcher must process the data to enable to analyse. Especially, stopword and typing error must be filter out in preprocessing to avoid affect to entire output. For this kind of pre preprocessing, two ways are often used. Make stopword dictionary or process by each word frequency [1].

However, it is different between Korean and English. There are few researches for Korean stopword dictionary. It is really hard to find Korean stopword list [2] So, in

this research, we suggest way to filter out the stopwords and typing error by its frequency and weight.

2 Related Research and Method

2.1 Existing Research

Gill-Hohyun(2018) extracted 10,000 of morpheme and removed important role words in sentence, formal morpheme with high frequency and meaninglessness. And Gill suggested draft proposal of Korean stopwords dictionary made by the words: high frequency and meaninglessness [3].

Lee-Minsik, Lee-Hongju (2016) devised Text-Word matrix to make division for customer review; is it worth it or not, to remove the wordes by scarcity and neutrality. And they classified the review in two group: the worth review group, and meaningless review group [4].

2.2 TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) is kind of weight model for information searching and text mining. It is statistical value which is judged each word is how important in each text file. It is used for extracting key words from documents, deciding web research key words ranking, and comparing similarity between documents.

TF (term frequency) is value of how often appear each words, DF(document frequency) is the value about the particular words is in the documents or not. IDF is inverse of DF. So, TF-IDF could be expressed by multiply of TF and IDF.

Table 1. TF-UDF Weight Model.

TF (term frequency)	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: *the count of words 'ti' in document 'dj'* $\sum_k n_{k,j}$: *the count of every appearance of every word in the documents 'dj'*
DF (document frequency)	$idf_i = \log \frac{\lvert D \rvert}{\lvert d_j \rvert t_i \in d_j \rvert}$ $\lvert D \rvert$: the number of documents belonged in documents set $\lvert d_j \rvert t_j \in d_j \rvert$: the number of documents which have the words t_j ,
TF-IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

3 Related Research and Method: Accumulative Frequency with TF-IDF Weight

3.1 Problem of TF-IDF

TF-IDF is multiplied value of TF and IDF. And it is expressed by the matrix: the number of 'unique' words * the number of documents file. For example, the number of words which is removed overlap is 4 and the number of documents (sentence) is 10, then as a result, the matrix has 4 * 10 size.

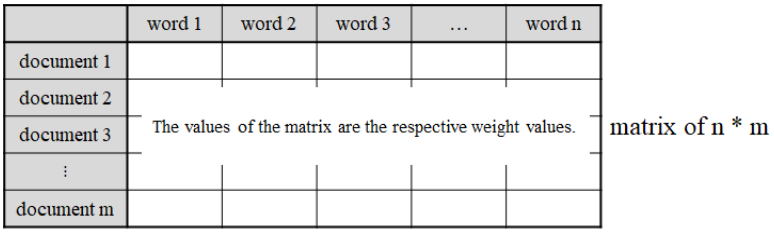

Fig. 1. Size of TF-IDF Result.

Such a small data like example, there are no problem. But the more words, the more data storage. The data storage needs are growth exponentially. For example, the example has 0.1millions of sentences and 1millions of words, it becomes 100,000 * 1,000,000 matrix and it means the number of 100,000 * 1,000,000 of TF-IDF weight.

3.2 Transformation of TF-IDF Result

In this research, the Hyper parameter(H) was decided to solve the exponential growth of TF-IDF value. In the documents, the number of the words more than particular number is 0, or it couldn`t, then the words get 1. It means, put every weight which size is m* n in the model and transform the weight between 0 or 1. After that, sum every result of the documents and make it as matrix which size is n* 2.

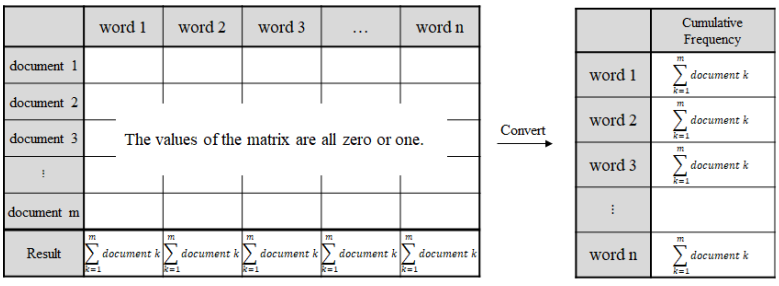

Fig. 2. Transform of TF-IDF and summate and Transform to n * 2.

In other words, TF-IDF weight value means importance of each words in each document. In this research, we decide criteria for hyperparameter H made by TF-IDF weight and classified ‘meaningful’ or ‘meaningless’ by H value. After that, summate every ‘meaningful’ word and calculated new accumulative frequency by TF-IDF weight.

3.2 Accumulative Frequency with Weight: Remove Stopwords

TF-IDF method gives low weight to meaningless words. Automatically, stopwords got low weight. It means, most of stopwords are excepted from accumulative frequency. It is possible to check the ranks without stopwords, typing error, meaningless words in this way.

4 CONCLUSION

5.1 Conclusion of Research

As we said, accumulative frequency by TF-IDF weight method is the way to classify meaningless words by hyper parameter(H) from TF-IDF weight. In existing research, researcher should create the stopwords dictionary first to remove the meaningless words. But this research has huge advantage that it is possible to remove the stopwords without other processing in very efficient way.

In this research, even we only used our method to only nouns, this method also can be used for verbs and adjectives. Especially, when doing sensitivity analysis by using verbs and adjectives, this method is good pre-process method to remove meaningless words.

5 REFERENCE

1. Miah K, Min S.: A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis. Korea Intelligent Information System Society. 18, 53--77 (2012)
2. Hohyun K.: The Study of Korean Stopwords list for Text mining. urimalgeulhakhoe. 78, 1--25 (2018)
3. Bongjun C. Hangjoo L.: A Generation and Matching Method of Normal-Transient Dictionary for Realtime Topic Detection. The Journal of KINGComputing. 13, 7--18 (2017)
4. Minsik L. Hongjoo L.: Increasing Accuracy of Classifying Useful Reviews by Removing Neutral Terms. Korea Intelligent Information System Society. 22, 129--142 (2016)