Korean Text Clustering With Syllable Vectors

Eon Young Park [1] , Bongsoo Jang [1]

1) *Department of Mathematical Sciences, Ulsan National Institute of Science and Technology(UNIST), Ulsan 44919, Republic of Korea*

Corresponding Author : Bongsoo Jang, bsjang@unist.ac.kr

ABSTRACT

Text mining, the process of deriving information from text, has been considered as one of most important technology in data sciences. Recently, many methods based on a neural network have been introduced to analyze information of text data. Doc2vec, one of those unsupervised techniques, creates a numeric representation of document to cluster documents. It is a small extension to Word2Vec. Instead using words to predict the next word, another feature vector, which is the topic of the paragraph, is added in a training process.

As increasing the number of documents, it is natural to increase the number of words so that the dimension of vector space of words is also getting larger. It loads a heavy training process in Doc2vec which induces the cost of computation. To overcome this difficulty, we propose PSV(preceding n-syllables vector) method to use pre-syllables rather than whole Korean word. Many Korean words come from Chinese characters so that each syllable has its meaning. It is motivated that preceding several syllables could have the similar meaning for the whole word. Since the words are represented by the combination of several syllables, the dimension of vector space of words is reduced and consistent. In this work, we present how well the proposed method has been performed compared with the Doc2vec for text clustering.

REFERENCES

1. LE, Quoc; MIKOLOV, Tomas. “Distributed representations of sentences and documents”. *In: International Conference on Machine Learning*. 2014. p. 1188-1196.

2. Kim, Kyunghoon. *A Mathematical Measurement For Korean Text Mining and Its Application*. 2018.