Embedding of compounds names from antioxidant articles is related to compound's features

Yuuto Matsumoto, YOKOHAMA National University

Recently, although many articles concerned with antioxidants have been published, we have not had an estimator of comprehensive antioxidant capacity because we cannot use all the information from the articles. Additionally, the antioxidant mechanism is complicated because it involves numerous factors that require explanation by a simple equation or correlation. Thus, a model that interprets all knowledge of articles in the view of chemistry is required. In this study, natural language processing was performed using Word2Vec on articles on antioxidant capacity. The representation vectors of compound names extracted from the documents were grouped into ten clusters. Our analysis of two clusters revealed that most of the compounds were flavonoids and flavonoid glycosides. A correlation between the descriptors and clusters was established by kernel density estimation and scatter plots of similarity, showing that the descriptors and clusters had a clear relationship. This confirms that a relationship exists between word vectors and compound descriptors using document analysis based on natural language processing.