Abstract

Semantic similarity is a task to estimate the strength of the semantic relationship between language units or concepts, in this case the common meaning possessed by the pair of words. The semantic similarity of the Indonesian words can be measured by using a knowledge base such as Big Indonesian Dictionary. The vector-based method is one of the methods for calculating semantic similarities. In this final project, I implement semantic similarity of Indonesian word pairsusing vector-based method, tf-idf weighting, and cosine similarity calculation.Using Big Indonesian Dictionary as knowledge base and golden standard based on SimLex999 and Rubenstein-goodenough consist of 180 pairs of words, we conduct the experiment.To evaluate the semantic similarity scores produced by the system, we ask 31 respondents to give a similarity score on the golden standard. We get the best correlation value of 0.5416 by adding definition of synonym in test. The best parameter that influences semantic equality value in this research is by adding the synonym without stopword removal.

Keywords: Big Indonesian Dictionary, cosine similarity, golden standard, semantic similarity, tf-idf, vector-based method.