Abstract

Textual Similarity/Relatedness can be the important base to relating between texts in another text mining fields, such as sentiment analysis, text categorization, etc. Those related texts can be similar/relate each other which set in lexical or semantic's way. So, the background of this research have issues that there is no available computer which can equate the human preception of semantic relatedness between words to ease in other text researches. That related texts can be calculates using the Explicit Semantic Analysis (ESA) method's. This method calculates the relationship between the two words that will be converted later into a score which ranged from 0-1 corresponding to the level of its relation. The database that used as reference in the calculation of these scores are from wikipedia articles that contain of 2000 titles, we use this because it was the largest source of knowledge based articles on the Internet. The testing began with compared two ESA tf.idf's calculation, biner and nonbiner. And to know how number of title articles influenced to score, the testing can be distinguished based on the number of the titles as many as 500, 1000 and 2000. After the scores has been established, then we compared and calculated the correlation with the gold standard wordsim353 relatedness. The correlation that generated form the system on 2000 titles with biner calculation is 0,2789, and nonbiner is 0,3958 for the wordsim353's gold standard. Whereas for the MEN's gold standard, The correlation that generated form the system on 2000 titles with biner calculation is 0,4249, and nonbiner is 0,4998.

Keywords : *Tekxtual Similarity, semantic, Explicit Semantic Analysis (ESA), score, word, tf.idf, wikipedia, gold standard , correlation, MEN, wordsim353 relatedness.*