INTISARI

DETEKSI SIMILARITAS DENGAN METODE *SEMANTIC* DAN *SYNTACTIC* DALAM PENULISAN TUGAS AKHIR

Oleh

SIDIQ TRI PRATIKTO
14/368540/PA/16292

Penelitian ini bertujuan untuk membangun program yang dapat mendeteksi tingkat similaritas dua dokumen tugas akhir yang berfokus kepada abstraknya. Terdapat banyak algoritma dalam mendeteksi tingkat similaritas dua dokumen di mana salah satunya menggunakan similaritas Jaccard (Jaccard, P., 1901). Proses perhitungan diawali dengan tahapan *pre-processing* abstrak-abstrak tugas akhir yang merupakan abstrak *suspected* dan yang merupakan abstrak sumber yang terdiri atas proses *text segmentation*, *stop word removal* dan *stemming*. Setelah masing-masing abstrak melalui tahapan *pre-processing*, selanjutkan setiap kalimat yang berada di abstrak *suspected* dipasangkan dengan setiap kalimat yang berada di abstrak sumber untuk dihitung nilai similaritasnya. Perhitungan dilakukan dengan menggunakan dua metode, yakni metode *semantic* dan metode *syntactic* yang kadarnya ditentukan suatu variabel α dalam rentang nilai $0 - 1$. Dalam proses perhitungan nilai similaritas pasangan kalimat dari abstrak *suspected* dan abstrak sumber, terdapat proses perhitungan nilai similaritas setiap dua kata di mana proses ini menggunakan konsep *word embeddings* dengan *data train* berupa dokumen-dokumen Skripsi Program Studi Ilmu Komputer UGM yang diimplementasikan menggunakan word2vec dan *vector cosine similarity*. Pasangan kalimat yang memiliki nilai similaritas melewati batas *threshold value* dinyatakan *similar* dan menjadi kandidat pasangan kalimat yang *similar* dalam pemeriksaan dua abstrak. Program ditulis dalam bahasa Python dan menggunakan pustaka Gensim yang di dalamnya terdapat word2vec. Hasil akhir dari penelitian ini memperlihatkan bahwa metode yang diterapkan pada penelitian ini memiliki hasil yang baik, yakni memiliki performa terbaik dengan nilai *Precision* didapatkan di angka 1, nilai *Recall* didapatkan di angka 1 dan *F-measure* didapatkan di angka 1, setara dengan metode similaritas Jaccard yang memiliki performa terbaik serupa. Namun, dikarenakan *dataset* yang tidak banyak, metode yang diterapkan di penelitian ini memberikan nilai similaritas yang cukup tinggi terhadap pasangan kalimat yang tidak *similar*, sehingga *threshold value* ditetapkan dengan nilai yang tinggi untuk mendapatkan performa yang baik.

Kata Kunci: Similaritas Jaccard, *semantic*, *syntactic*, pustaka Gensim, *pre-processing*, *text segmentation*, *stop word removal*, *stemming*, *suspected*, sumber, *word*

embeddings, *vector cosine similarity*, *threshold value*, *precision*, *recall*, *F-measure*

ABSTRACT

SIMILARITY DETECTION WITH SEMANTIC AND SYNTACTIC METHOD ON THESIS WRITING

By

SIDIQ TRI PRATIKTO
14/368540/PA/16292

This research aims to build a program that can detect the level of similarity of two thesis that focused on the abstract. There are many algorithms in detecting the level of similarity of two documents, one of which uses Jaccard similarity (Jaccard, P., 1901). The calculation process begins with the *pre-processing* abstract thesis which is the *suspected* abstract and which is the source abstract consisting of *text segmentation*, *stop word removal* and *stemming*. After each abstract goes through the *pre-processing* step, then each sentences in the abstract *suspected* is paired with every sentences in the source abstract to calculate the similarity value. The calculation is carried out using two methods, namely the *semantic* method and the *syntactic* method whose content is determined by a variable α in the value range of $0 - 1$. In the process of calculating the similarity value of sentence pairs from the *suspected* abstract and the source abstract, there is a process of calculating the similarity value of each two words where this process uses the concept of *word embeddings* with *data train* in the form of Computer Science UGM thesis documents implemented using word2vec and *vector cosine similarity*. Sentence pairs that have similarity values exceeding the *threshold value* are declared *similar* and become candidate pairs of sentences that are *similar* in the examination of two abstracts. The program is written in Python and uses the Gensim library which includes word2vec. The final result of this study shows that the method applied in this study has good results, namely it has the best performance with the value of *Precision* obtained at number 1, the value of *Recall* obtained at number 1 and *F- measure* obtained at number 1, equivalent to the Jaccard similarity method which has the same best performance result. However, because there are not many *dataset*, the method applied in this study gives a fairly high similarity value to sentence pairs that are not *similar*, so *threshold value* is set with a high value to get performance the good one.

Keywords: Jaccard similarity, *semantic*, *syntactic*, Gensim library, *pre-processing*, *text segmentation*, *stop word removal*, *stemming*, *suspected*, source, *word embeddings*, *vector cosine similarity*, *threshold value*, *precision*, *recall*, *F -measures*