INTISARI

Pendeteksian Kemiripan Sinopsis Rencana Penelitian Tugas Akhir dan Skripsi Menggunakan Topic2Vec
(Studi Kasus di STMIK Bumigora Mataram)

Oleh

Zaenal Abidin
15/388515/PPA/04954

Pencarian kemiripan dokumen merupakan hal yang sudah umum dilakukan oleh para peneliti. Banyak pendekatan model yang sudah dikembangkan, diantaranya adalah model probabilitas dan prediksi. Model probabilitas mencari kemiripan dokumen berdasarkan kemunculan kata pada dokumen sedangkan model prediksi dengan mempelajari kotek kata yang ada pada dokumen dan direpresentasikan dalam model *word embedding*. *Word embedding* yang dihasilkan model prediksi menjadi acuan dalam pencarian kemiripan pada dokumen.

Pada penelitian ini, dilakukan pencarian kemiripan sinopsis penelitian tugas akhir dan skripsi untuk menilai kelayakan penyusunan tugas akhir dan skripsi di STMIK Bumigora Mataram. Pencarian kemiripan dokumen mengkombinasikan model probabilitas dengan model prediksi. Salah satu metode yang mengkombinasikan model tersebut adalah Topic2Vec. Topic2Vec membangkitkan topik dari kumpulan tugas akhir dan skripsi dengan model probabilitas, *latent dirichlet allocation* (LDA). Hasil topik yang dibangkitkan dari model probabilitas dibentuk ke *word embedding* dengan model prediksi (skip-gram). *Word embedding* yang dihasilkan, digunakan untuk mencari kemiripan dokumen menggunakan metode *word mover distance*.

Pengujian akurasi yang dilakukan dari model dengan menggunakan *mean average precision* (MAP) dan *recall*, menunjukkan bahwa model Topic2Vec tidak mampu meningkatkan akurasi untuk mencari kemiripan dokumen. Pengujian menunjukkan hasil dengan MAP 56% dan *recall* 64% pada jumlah 1300 topik yang dibangkitkan.

Kata kunci: kemiripan dokumen, LDA, *word embedding*, *skip-gram*, Topic2Vec.

ABSTRACT

Detecting Similarities of Synopses of Research Plan of Final Projects and Undergraduate Theses Using Topic2Vec (Case Study in STMIK Bumigora Mataram)

by

Zaenal Abidin
15/388515/PPA/04954

Document similarity search is commonly done by researchers. Many approach models have been developed, such as probability and prediction models. The probability model search document similarities based on word appearance in documents while the prediction model trains the existing word contexts in the documents to generated word embedding. Base on the word embedding is used to find documents similarity.

This research is conducted to find similarities of synopsis of research final project and thesis in STMIK Bumigora Mataram to assess the feasibility of the preparation of the final project and thesis. Document similarity search combines probability model and prediction model. The method that combines the model is Topic2Vec. Topic2Vec generates topics from the final project and theses using the probability model, latent dirichelet allocation (LDA). The generated topic results are formed into word embedding using the prediction model (skip-gram). Generated Word embedding is used to find document similarity using word mover distance method.

After conducting accuracy testing by applying the model and employing the mean average precision (MAP) and recall, figured out that the Topic2Vec model was unable to improve accuracy to find similarities among documents. The testing presented MAP of 56% and recall of 64% from 1300 topics raised.

Keywords: document similarity, LDA, word embedding, skip-gram, Topic2Vec.