INTISARI

Analisis Perbandingan Metode Jaccard Coefficient dan Cosine Similarity untuk Kemiripan Teks Bahasa Indonesia

Oleh

Sheraton Pawestri

21/486087/PPA/06234

Informasi yang cukup mudah didapat mempunyai banyak manfaat. Manfaat dengan membuat sebuah model untuk mendeteksi kemiripan (*similarity*) antar dokumen misal untuk cek plagiasi, cek jawaban *essay* siswa, pengelompokan teks berdasarkan topik, pengelompokan judul penelitian berdasarkan tema, dan masih banyak lagi. Banyaknya manfaat menyebabkan penelitian mengenai cek kemiripan layak untuk dikembangkan.

Penelitian cek kemiripan untuk bahasa Indonesia masih sedikit. Sehingga penelitian ini dilakukan analisis komparasi kinerja Doc2Vec dengan Jaccard Coefficient dan Cosine Similarity untuk cek kemiripan antar dokumen Bahasa Indonesia berbagai topik (Bisnis, Hiburan, Teknologi, dan Kesehatan). *Dataset* yang digunakan sebanyak 3 dataset untuk dapat dianalisis satu persatu. *Dataset* pertama dari GoogleNews sebanyak 200 berita. *Dataset* kedua menggunakan *dataset* IndoNLU (*benchmark*) sebanyak 300 data. *Dataset* ketiga menggunakan Tappaco sebanyak 1602 data (*benchmark*). Penelitian ini diharapkan mampu menganalisis perbandingan kinerja tiap *dataset* yang digunakan.

Hasil penelitian ini menunjukkan Cosine Similarity lebih unggul dibandingkan dengan Jaccard Coefficient dalam hal akurasi, presisi, *recall*, dan nilai f-1. Sementara *dataset* yang paling unggul kinerjanya adalah Google News.

Kata Kunci: teks berita bahasa Indonesia, cek kemiripan, doc2vec, Jaccard Coefficient, Cosine Similarity

ABSTRACT

Comparative Analysis of Jaccard Coefficient and Cosine Similarity Methods for Indonesian Language Text Similarities

By

Sheraton Pawestri

21/486087/PPA/06234

Information that is easy to obtain has many benefits. The benefits of similarities between documents are checking plagiarism, checking student essay answers, grouping texts based on topics, grouping research titles based on themes, and many more. These benefits make research on similarity checks worth developing.

Similarity check research for Indonesian is still scanty. So, this study conducted a comparative analysis of the performance of Doc2Vec with the Jaccard Coefficient and Cosine Similarity to check similarities between Indonesian language documents on various topics (Business, Entertainment, Technology, and Health). The datasets used are three datasets to be analyzed one by one. First, dataset from Google News with 200 news. Second, dataset uses the IndoNLU dataset of 300 data. The third dataset uses the TaPaCo benchmark of 1602 data. This research's goal is to be able to analyze the performance comparison of each dataset used.

The results of this study show that Cosine Similarity is superior to Jaccard Coefficient in terms of accuracy, precision, recall, and f-1 value. The dataset with the best performance is Google News.

Keywords: Indonesian news text, similarity, doc2vec, Jaccard Coefficient, Cosine Similarity