ABSTRAK

IDENTIFIKASI KEMIRIPAN DOKUMEN AKADEMIK PADA TUGAS KULIAH MAHASISWA DI JURUSAN ILMU KOMPUTER UNIVERSITAS LAMPUNG

Oleh

THERESIA TRI OKTAVIA IRMAWANTI

Integritas akademik dalam lingkungan perguruan tinggi, terutama pada program studi ilmu komputer, sangat bergantung pada kemampuan mendeteksi kemiripan antar dokumen yang dikumpulkan mahasiswa, baik berupa laporan tertulis maupun kode program. Penelitian ini merancang dan membangun sebuah sistem deteksi kemiripan dokumen akademik yang mampu memproses dokumen teks berbahasa Indonesia dan kode sumber Python secara terintegrasi dalam satu platform. Pada tahap pemrosesan dokumen teks, digunakan kombinasi representasi berbasis statistik TF-IDF dan representasi berbasis semantik Word2Vec (FastText), dengan bobot 0,6 untuk TF-IDF dan 0,4 untuk FastText, yang selanjutnya dihitung kemiripannya menggunakan metode *Cosine Similarity*. Pendekatan ini dirancang agar sistem tidak hanya mampu menangkap kesamaan kata secara langsung, tetapi juga mengenali kemiripan makna meskipun kalimat telah mengalami parafrase. Sementara itu, pada pemrosesan kesamaan kode program, sistem menggunakan strategi penggabungan dua skor kemiripan, yaitu TF-IDF *Cosine Similarity* dan *Line-based Cosine Similarity*, untuk menghasilkan penilaian yang lebih akurat dan menyeluruh. Guna mengatasi tantangan efisiensi pada dataset berskala besar, algoritma *K-Means Clustering* diimplementasikan untuk mengelompokkan dokumen sehingga proses perbandingan hanya dilakukan di dalam klaster yang relevan. Hasil pengujian terhadap 90 dokumen teks dan 117 file kode Python memperlihatkan bahwa sistem mampu mendeteksi masing-masing 6 pasang dokumen dan kode yang memiliki kemiripan tinggi. Selain itu, penggunaan klasterisasi terbukti menekan jumlah komputasi perbandingan secara signifikan, yakni sebesar 85,8% untuk dokumen teks dan 83,7% untuk kode sumber, dengan peningkatan kecepatan proses mencapai 1,90x dan 6,86x. Sistem ini diimplementasikan dalam bentuk aplikasi web menggunakan framework Flask, yang menyediakan modul pemrosesan dokumen teks (SimDoc) dan kode sumber (SimCode), visualisasi bagian yang mirip serta fitur ekspor hasil analisis ke dalam format laporan PDF.

Kata Kunci: deteksi kemiripan dokumen; *cosine similarity*; *k-means clustering*; TF-IDF; *word2vec*.

ABSTRACT

IDENTIFICATION OF DOCUMENT SIMILARITIES IN ACADEMIC ASSIGNMENTS SUBMITTED BY STUDENTS IN THE DEPARTMENT OF COMPUTER SCIENCE

By

THERESIA TRI OKTAVIA IRMAWANTI

Academic integrity in higher education, particularly in computer science programs, relies heavily on the ability to detect similarities among documents submitted by students, whether in the form of written reports or programming code. This research designs and develops an academic document similarity detection system capable of processing Indonesian text documents and Python source code in an integrated manner within a single platform. For text document processing, a combination of statistical-based representation using TF-IDF and semantic-based representation using Word2Vec (FastText) is employed, with weights of 0.6 for TF-IDF and 0.4 for FastText, and similarity is subsequently measured using the Cosine Similarity method. This approach is designed so that the system is not only able to capture direct word-level matches, but also recognize semantic similarities even when sentences have been paraphrased. Meanwhile, for source code similarity detection, the system applies a dual-score fusion strategy combining TF-IDF Cosine Similarity and Line-based Cosine Similarity to produce a more accurate and comprehensive similarity assessment. To address computational efficiency challenges on large-scale datasets, the K-Means Clustering algorithm is implemented to group documents so that pairwise comparisons are performed only within relevant clusters. Experimental results on 90 text documents and 117 Python source code files demonstrate that the system successfully detected 6 pairs of similar documents and 6 pairs of similar source code with high similarity scores. Furthermore, the use of clustering significantly reduced the number of pairwise comparisons by 85.8% for text documents and 83.7% for source code, with processing speed improvements of 1.90x and 6.86x, respectively. The system is implemented as a Flask-based web application featuring a text document processing module (SimDoc) and a source code processing module (SimCode), along with similarity visualization and the ability to export analysis results as PDF reports.

Keywords: *document similarity detection; cosine similarity; k-means clustering; TF-IDF; word2vec.*