INTISARI

MENEMUKAN SIMILARITAS DALAM DOKUMEN MENGGUNAKAN TF-IDF, SENTENCE TRANSFORMER, DAN DISTILBERT SEBAGAI VEKTORISASI DAN SIMILARITAS KOSINUS SEBAGAI METODE PERHITUNGAN

By

Raphael Octavian Bong

18/425532/PA/18424

Tesis ini mengeksplorasi tema kesamaan dokumen melalui pemanfaatan teknik vektorisasi dan metode perhitungan, dengan menggunakan Python sebagai bahasa pemrogramannya. Penekanan utamanya adalah menyelidiki kemanjuran TF-IDF, Sentence transformer, dan DistilBERT sebagai metode vektorisasi, ditambah dengan penerapan kesamaan kosinus untuk perhitungan.

Penelitian dimulai dengan pemilihan beragam kumpulan data yang terdiri dari makalah penelitian yang berfokus pada Natural Language Processing (NLP). Pemrosesan awal teks kemudian digunakan untuk mengoptimalkan hasil vektorisasi. Langkah terakhir melibatkan penggunaan kesamaan kosinus sebagai metode perhitungan untuk mengukur kesamaan antar dokumen. Studi ini bertujuan untuk menyumbangkan wawasan berharga bagi para peneliti yang mencari inspirasi untuk menyempurnakan makalah mereka atau merintis inovasi baru.

Dengan menggunakan berbagai teknik vektorisasi, termasuk TF-IDF, Sentence transformer, dan DistilBERT, dengan metode penghitungan kesamaan kosinus yang konsisten, hasilnya mengungkapkan perbedaan yang patut diperhatikan. TF-IDF menunjukkan rata-rata kesamaan yang lebih rendah yaitu 0,07%, sedangkan Sentence transformer dan DistilBERT menunjukkan rata-rata masing-masing 81,53% dan 91,32%. Perbedaan ini menunjukkan bahwa metode vektorisasi berbasis transformator mungkin menawarkan pendekatan yang lebih berbeda dalam mendeteksi kesamaan.

Keta Kunci: Python, persamaan, pra-pemrosesan teks, TF-IDF, Sentence transformer, DistilBERT, Kemiripan Kosinus.

ABSTRACT

FINDING SIMILARITY WITHIN DOCUMENTS USING TF-IDF, SENTENCE TRANSFORMER, AND DISTILBERT AS VECTORIZATION AND COSINE SIMILARITY AS CALCULATION METHOD.

By

Raphael Octavian Bong

18/425532/PA/18424

This thesis explores the theme of document similarity through the utilization of vectorization techniques and calculation methods, employing Python as the programming language. The primary emphasis is on investigating the efficacy of TF-IDF, Sentence transformer, and DistilBERT as vectorization methods, coupled with the application of cosine similarity for calculations.

The research commences with the selection of a diverse dataset comprising research papers focused on Natural Language Processing (NLP). Text preprocessing is then employed to optimize the results of vectorization. The final step involves the use of cosine similarity as the calculation method to quantify the similarity between documents. This study aims to contribute valuable insights for researchers seeking inspiration to enhance their papers or pioneer new innovations.

By employing various vectorization techniques, including TF-IDF, Sentence transformer, and DistilBERT, with a consistent calculation method of cosine similarity, the results reveal noteworthy distinctions. TF-IDF exhibits a lower average similarity of 0.07%, while Sentence transformer and DistilBERT show averages of 81.53% and 91.32%, respectively. This disparity suggests that transformer-based vectorization methods may offer a more nuanced approach to similarity detection.

Keywords: Python, pre-process text, similarity, TF-IDF, Sentence transformer, DistilBERT, Cosine Similarity.