Universitas Indonesia Library >> UI - Skripsi Membership

Studi Perbandingan Metode Clustering K-Means, DBSCAN, dan HDBSCAN pada BERTopic untuk Pendeteksian Topik = Comparative Study of K-Means, DBSCAN, and HDBSCAN Clustering Methods on BERTopic for Topic Detection

Julizar Isya Pandu Wangsa, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=9999920531348&lokasi=lokal

---

Abstrak

Pendeteksian topik merupakan suatu proses pengidentifikasian suatu tema sentral yang ada dalam kumpulan dokumen yang luas dan tidak terorganisir. Hal ini merupakan hal sederhana yang bisa dilakukan secara manual jika data yang ada hanya sedikit. Untuk data yang banyak dibutuhkan pengolahan yang tepat agar representasi topik dari setiap dokumen didapat dengan cepat dan akurat sehingga machine learning diperlukan. BERTopic adalah metode pemodelan topik yang memanfaatkan teknik clustering dengan menggunakan model pre-trained Bidirectional Encoder Representations from Transformers (BERT) untuk melakukan representasi teks dan Class based Term Frequency Invers Document Frequency (c-TF-IDF) untuk ekstraksi topik. Metode clustering yang digunakan pada penelitian ini adalah metode -K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), dan Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). BERT dipilih sebagai metode representasi teks pada penelitian ini karena BERT merepresentasikan suatu kalimat berdasarkan sequence-of-word dan telah memperhatikan aspek kontekstual kata tersebut dalam kalimat. Hasil representasi teks merupakan vektor numerik dengan dimensi yang besar sehingga perlu dilakukan reduksi dimensi menggunakan Uniform Manifold Approximation and Projection (UMAP) sebelum clustering dilakukan. Model BERTopic dengan tiga metode clustering ini akan dianalisis kinerjanya berdasarkan matrik nilai coherence, diversity, dan quality score. Nilai quality score merupakan perkalian dari nilai coherence dengan nilai diversity. Hasil simulasi yang didapat adalah model BERTopic menggunakan metode clustering K-Means lebih unggul 2 dari 3 dataset untuk nilai quality score dari kedua metode clustering yang ada.
.....Topic detection is the process of identifying a central theme in a large, unorganized collection of documents. This is a simple thing that can be done manually if there is only a small amount of data. For large amounts of data, proper processing is needed to represent the topic of each document quickly and accurately, so machine learning is required. BERTopic is a topic modeling method that utilizes clustering techniques by using pre-trained Bidirectional Encoder Representations from Transformers (BERT) models to perform text representation and Class based Term Frequency Inverse Document Frequency (c-TF-IDF) for topic extraction. The clustering methods used in this research are the K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). BERT was chosen as the text representation method in this research because BERT represents a sentence based on sequence-of-words and has considered the contextual aspects of the word in the sentence. The result of text representation is a numeric vector with large dimensions, so it is necessary to reduce the dimensions using Uniform Manifold Approximation and Projection (UMAP) before clustering is done. The BERTopic model with three clustering methods will be analyzed for performance based on the matrix of coherence, diversity, and quality score values. The quality score value is the multiplication of the coherence value with the diversity value. The simulation results obtained are the

BERTopic model using K-Means clustering method is superior to 2 of the 3 datasets for the quality score value of the two existing clustering methods.