INTISARI

ANALISIS N-GRAM MENGGUNAKAN VECTOR SPACE MODEL PADA TEXT SIMILARITY

Diusulkan oleh
MUHAMMAD HIDAYATULLOH
16/403695/PPA/05212

Salah satu metode ektraksi fitur dengan konsep *Bag of Wor*d (BOW) yang populerr digunakan adalah *term frequency-inverse document frequency* (TF-IDF). TF-IDF merupakan ekstraksi fitur yang tidak memperhatikan urutan kata sebuah kalimat sehingga ektraksi fitur yang dihasilkan berdasarkan pemenggalan setiap term/kata dalam sebuah kalimat. Jika terdapat lebih dari satu kalimat yang berbeda dengan komposisi kata yang sama, maka kalimat tersebut dianggap mirip dan jika sebuah kalimat terdapat 2 kata kunci atau lebih maka ektraksi fitur yang dihasilkan berdasarkan *term* tidak relevan hal tersebut akan mempengaruhi akurasi.

Penelitian ini berjuan untuk memperbaiki kelemahan *term* tunggal pada tahap ektraksi fitur diperbaiki dalam penelitian ini dengan menambahkan metode *n-gram. N-gram* digunakan untuk memotong kata pada suatu kalimat berdasarkan jumlah n-kata. Metode pendukung untuk mengetahui performa *n-gram* digunakan *Vector space model* dengan pembobotan TF-IDF dan perhitungan similaritas menggunakan *cosine similarity*.

Berdasarkan skenario pengujian yang dilakukan, pengujian performa akurasi dengan membandingkan 1-gram sampai 1-7-gram menggunakan 244 data penyakit infeksi dan 120 *query* pengujian penyakit dihasilkan performa tingkat akurasi terbaik menggunakan skenario *full preprocessing* dengan menerapkan 1-3-*gram* dan hasil akurasi yang didapatkan adalah 91.7%.

Kata kunci – *preprocessing, n-gram, cosine similarity, vector space model*

ABSTRACT

ANALYSIS OF N-GRAM USING VECTOR SPACE MODEL IN TEXT SIMILARITY

by

MUHAMMAD HIDAYATULLOH
16/403695/PPA/05212

The process of text similarity is to compare document similarities to be used as an information retrieval tool. The ability to search for similar documents is usually implemented in text data to be extracted information. Feature extraction plays a role in determining which features will be used by the text-similarity technique. If the value of the resulting feature is incorrect, then the information extracted cannot fulfill the desired criteria. As a result, the processing of documents that you want to compare does not meet the user's wishes.

One of the feature extraction methods with the popular and popular BOW concept is the term frequency-inverse document frequency (TF-IDF). TF-IDF is a feature extraction that does not pay attention to the word order of a sentence and feature extraction that is generated based on the decapitation of terms/words in a sentence. If there is more than one different sentence with the same word composition, then the sentence will be considered similar and if a sentence contains 2 or more keywords then the extraction of features generated based on irrelevant terms will affect accuracy. To improve the deficiencies, this study uses a support method for feature extraction, namely n-gram.

Based on the test results, the best accuracy level performance is generated from a full preprocessing scenario by applying a 1-2-gram and the accuracy results obtained is 96%.

Keywords: preprocessing, n-gram, cosine similarity, accuracy, vector space model