Penggunaan Word Embedding dan Bobot Kata pada Algoritma Textrank untuk Peringkasan Artikel Bahasa Indonesia = The Use of Word Embedding and Word Weight in Textrank Algorithm for Summarizing Indonesian Articles

Nicholas Pangestu, author

Abstrak

Panjangnya suatu berita terkadang mengurangi minat seseorang untuk membaca berita, hal ini dapat kita lihat dari banyaknya istilah “tl:dr” pada thread di internet. Peringkasan dokumen dapat menciptkan ringkasan berita dan mengurangi waktu yang dibutuhkan untuk membaca. Salah satu cara yang dapat digunakan untuk melakukan peringkasan dokumen adalah menggunakan algoritma Textrank. Pada penelitian ini akan diimplementasikan word embedding untuk membantu algoritma Textrank memahami makna suatu kata dengan lebih baik. Hasil yang didapatkan menunjukkan bahwa penggunaan word embedding meningkatkan performa dari algoritma Textrank hingga 13% pada ROUGE-1 dan hingga 21% pada ROUGE-2. Model word embedding BERT memiliki performa tertinggi jika dibandingkan dengan word2vec (3% lebih tinggi pada ROUGE-1 dan 7% lebih tinggi pada ROUGE-2) dan fasttext (5% lebih tinggi pada ROUGE-1 dan 10% lebih tinggi pada ROUGE-2). Pada penelitian ini juga mengimplementasikan pembobotan TF-IDF dalam membuat sebuah representasi suatu kata. Hasil yang didapatkan menunjukkan bahwa pembobotan TF-IDF dapat meningkatkan performa dari tiap model word embedding yang digunakan hingga 11% pada ROUGE-1 dan hingga 19% pada ROUGE-2 dibandingkan performa tanpa pembobotan TF-IDF.

.....The length of article news sometimes reduces one's interest in reading the news, we can see this from the many terms "tl:dr" in threads on the internet. Document summarization can create news summaries and reduce the time it takes to read. One way to do document summarization is to use the Textrank algorithm. In this research, word embedding will be implemented to help the Textrank algorithm understand the meaning of a word better. The results show that the use of word embedding improves the performance of the Textrank algorithm up to 13% in ROUGE-1 and up to 21% in ROUGE-2. BERT word embedding model has the highest performance when compared to word2vec (3% higher in ROUGE-1 and 7% higher in ROUGE-2) and fasttext (5% higher in ROUGE-1 and 10% higher in ROUGE-2). This study also implements TF-IDF weighting to make a word representation. The results show that TF-IDF weighting can improve the performance of each word embedding model used up to 11% in ROUGE-1 and 19% in ROUGE-2 compared to the performance without using TF-IDF.