Segmentasi dokumen teks berbahasa Indonesia menggunakan metode text tiling

Siahaan, Edison Pardengganan, author

Abstrak

Penelitian yang dilakukan pada tesis ini dimotivasi oleh adanya kebutuhan untuk dapat melakukan pengelolaan informasi pada dokumen suara khususnya berita berbahasa Indonesia. Informasi pada dokumen suara berita berbahasa Indonesia dapat diubah menjadi informasi berbentuk dokumen teks, dengan menggunakan perangkat lunak Automatic Speech Recognition (ASR). Pada penelitian ini perangkat ASR yang digunakan adalah perangkat ASR Sphinx 4.

Penggunaan perangkat Sphinx 4 ini didasari telah dilakukannya penelitian tentang transkripsi dokumen suara berbahasa Indonesia menggunakan perangkat ini. Hasil keluaran dari ASR berupa dokumen teks yang tidak memiliki batasan akhir dan tidak tersegmentasi secara jelas, tentu menyulitkan dalam pengolahan data teks tersebut. Dalam kerangka itu, maka penelitian yang dilakukan pada tesis ini ditujukan untuk mengetahui metode yang efektif dalam melakukan segmentasi hasil transkripsi berita suara berbahasa Indonesia. Metode yang akan diuji pada penelitian ini adalah metode TextTiling berbasis perbandingan blok dengan pembobotan TF-IDF-Mutual Information, TF-IDFMutual Information-Word Similarity, TF-IDF-Word Frequency, TF-IDF, Latent Semantic Analysis dan metode TextTiling berbasis Vocabulary Introduction. Segmentasi dilakukan untuk berita teks dan dokumen teks hasil transkripsi berita suara yang telahdikatagorikan menjadi 5 topik yaitu topik politik, sosial budaya, ekonomi, hukum dan olah raga. Hasil pengujian terhadap masing-masing teknik pembobotan menunjukkan bahwa metode segmentasi TextTiling dengan teknik pembobotan TF-IDF-Word Frequency merupakan metode segmentasi yang paling baik untuk dipakai dalam melakukan segmentasi hasil transkripsi dari perangkat pengenal suara (Automatic Speech Recognition). Pada penelitian ini telah dibuktikan bahwa teknik pembobotan TF-IDF-Word Frequency memiliki ketepatan segmentasi lebih tinggi baik pada dokumen teks hasil transkripsi (81,4%) ataupun pada dokumen berita teks (73,3%). Metode segmentasi yang dilakukan pada penelitian ini dapat terus dikembangkan menggunakan teknik-teknik lain dalam menunjang proses segmentasi hasil transkripsi berita berberbahasa Indonesia, seperti mempergunakan metode-metode optimalisasi dalam memperoleh urutan batas segmen yang optimal.