Pengembangan pemodelan topik bahasa indonesia memanfaatkan latent dirichlet allocation = Development of indonesian language topic modeling using latent dirichlet allocation

Arman, author

Abstrak

ABSTRAK
Ekstraksi topik merupakan tugas utama dalam penambangan teks sebagai upaya mengeluarkan informasi yang terpendam dalam teks secara heuristik. Proses ini dilakukan lewat pemodelan topik yakni sebuah proses mengidentifikasi topik- topik yang ada dalam sebuah objek teks atau menurunkan pola-pola tersembunyi dalam sebuah korpus teks. Dalam penelitian ini pemodelan topik diaplikasikan pada data teks berbahasa Indonesia menggunakan modul program bernama Gensim dalam bahasa pemrograman Python. Dataset terdiri dari 93 dokumen berita daring Kompas dengan beragam klasifikasi. Jumlah topik optimal yang diperoleh diuji menggunakan machine learning clustering k-means. Dalam proses penelitian ini ternyata diperlukan suatu mekanisma umpanbalik manual untuk mereduksi noise agar diperoleh pemodelan topik yang lebih baik. Hasil uji memperlihatkan teknik Latent Dirichlet Allocation LDA yang telah ditingkatkan / dimodifikasi LDA as LSI memiliki koherensi topik yang jauh lebih baik dibanding teknik LDA saja dalam penelitian ini: 0.94 dibanding 0.34 . Koherensi yang tinggi mengindikasikan bahwa topik hasil pemodelan ini merupakan topik yang dapat dijelaskan dengan sedikit label.
<hr />
ABSTRACT
Topic extraction is main task in text mining as an effort to dig buried information within text heuristically. This process is done through topic modeling, a process to identify topics within text object or to derive hidden patterns in a text corpus. In this research, topic modeling is applied to Indonesian language texts using Gensim module in Python programming language. The dataset consists of 93 online news documents from Indonesian national newspaper, Kompas, with several different classifications. The identified optimum number of topics k is visualized using clustering machine learning k means. In the process of this research turned out to need a mechanism of manual feedback for noise reduction in order to get better topic modeling. The test results show that enhanced modified Latent Dirichlet Allocation LDA as LSI has a much better topic coherence than LDA technique alone in this study 0.94 compared to 0.34 . High coherence indicates that topics resulting from this topic modeling is a topic that can be explained with few labels.