

ABSTRAK

Beragamnya dokumen skripsi dengan topik dan metode yang beragam pula, menjadikan tantangan mahasiswa dalam menyusun skripsi. Topik dan metode yang diteliti pada penelitian sebelumnya sangat beragam menjadi dilema mahasiswa dalam menentukan topik dan metode yang akan diteliti. Adanya masalah tersebut peneliti ingin melakukan *topic modeling* dari abstrak dokumen skripsi untuk menganalisis topik dan metode dokumen skripsi yang ada.

Salah satu metode *topic modeling* adalah metode *Latent Dirichlet Allocation*. Abstrak dokumen skripsi akan digunakan sebagai objek pada penelitian ini karena abstrak merupakan ringkasan dari keseluruhan isi karya tulis ilmiah yang mengandung inti dari pembahasan penelitian yang dilakukan. Pengumpulan data didapat dari *website* Eprints UPN "Veteran" Yogyakarta dengan alamat *url* <http://eprints.upnyk.ac.id/>. Data diambil dengan metode *web scraping*. Data berisi *code_subject*, *subject*, *title*, *author*, dan *abstract*. Jumlah data yang didapat dari *web scraping* sebanyak 35.650 data.

Hasil nilai *coherence score* dari model LDA yang dikembangkan untuk pemodelan topik jurnal tertinggi dari tiap kategori adalah 0.7045 pada kategori *Agriculture (General)* dengan jumlah topik sebanyak enam dan nilai terendah adalah 0.3788 pada kategori *Computer software* dengan jumlah topik sebanyak empat. Nilai rata-rata dari *coherence score* dari semua kategori didapat sebesar 0.5246 dan rata-rata jumlah topik sebanyak enam. Hasil pengujian dari sepuluh sampel yang digunakan sebagai dokumen uji didapatkan hasil topik yang tidak relevan sebanyak empat dan topik yang relevan sebanyak enam dengan nilai akurasi sebesar 60%. Berdasarkan hasil tersebut, dapat disimpulkan bahwa model LDA yang dikembangkan belum mampu menghasilkan topik yang relevan dengan dokumen jurnal yang ada.

Kata kunci: *text mining; topic modeling; latent dirichlet allocation; gibbs sampling.*

ABSTRACT

The variety of thesis documents with various topics and methods also makes it a challenge for students to write a thesis. The topics and methods studied in previous research were very diverse, creating a dilemma for students in determining the topics and methods to be researched. The existence of this problem researchers wants to do topic modeling from the abstract of the thesis document to analyze the topics and methods of the existing thesis document.

One method topic modeling is a method Latent Dirichlet Allocation. The abstract of the thesis document will be used as the object of this research because the abstract is a summary of the entire content of a scientific paper which contains the essence of the discussion of the research carried out. Data collection was obtained from the website UPN "Veteran" Yogyakarta eprints with url <http://eprints.upnyk.ac.id/>. Data was taken using the method web scraping. Data contains code_subject, subject, title, author, and abstract. The amount of data obtained from web scraping is 35,650 data.

Value result of coherence score from the LDA model developed for topic modeling for journal topics, the highest for each category is 0.7045 in the category Agriculture (General) with a total of six topics and the lowest score is 0.3788 in the category Computer software with a total of four topics. Average value of coherence score from all categories was obtained at 0.5246 and the average number of topics was six. The test results of the ten samples used as test documents showed four irrelevant topics and six relevant topics with an accuracy value of 60%. Based on these results, it can be concluded that the LDA model developed has not been able to produce topics that are relevant to existing journal documents.

Keywords: text mining; topic modeling; latent dirichlet allocation; gibbs sampling.