Jurnal "LOG!K@" , Jilid 7, No. 1, 2017, Hal. 46 - 51
ISSN 1978 – 8568

APLIKASI LATENT DIRICHLET ALLOCATION (LDA) PADA *CLUSTERING* DATA TEKS

Zulhanif, Sudartianto, Bertho Tantular, I Gede Nyoman Mindra Jaya
Departemen Statistika FMIPA
Universitas Padjadjaran

Abstract: Latent Dirichlet Allocation (LDA), generative probabilistic model on a set of text data (corpus). LDA is a Bayesian model of hierarchy, where a bunch of text is modeled as a model mixture of different topics. In the context of text modeling, probability of topics gives an explicit representation of a document. In this study presents the inference techniques based Gibbs Sampling algorithm for estimating parameters of Bayesian modeling and classification of text documents.
Keywords: *Text Mining, LDA, Gibbs Sampling Algorithm*.

Abstrak: *Latent Dirichlet Allocation* (LDA), model probabilistik generatif pada sekumpulan data teks (*corpus*) . LDA adalah model Bayesian Hirarki , di mana sekumpulan teks dimodelkan sebagai model campuran dari berbagai topik. Dalam kontek pemodelan teks, probabilitas topik memberikan representasi eksplisit dari sebuah dokumen. Pada penelitian ini menyajikan teknik inferensi berdasarkan algoritma *Gibbs Sampling* untuk mengestimasi parameter Bayes dalam pemodelan dokumen dan klasifikasi teks.
Kata kunci: *Text Mining, LDA, Gibbs Sampling Algorithm*.

PENDAHULUAN

Perkembangan analisis teks sendiri pada dasanya bersumber pada matriks *term frequency–inverse document frequency (tf-idf*), (Salton and McGill, 1983). Selanjutnya berlanjut pada perkembangan pereduksian mariks *tf-idf* dengan menggunakan metode pereduksian dimensi seperti *Latent Semantic Analysis* (LSA) dan *Probabilistic Latent Semantic Analysis* (PLSA). *Latent Semantic Analysis* (LSA) metode yang dipatenkan pada tahun 1988 (US Patent 4,839,853) oleh Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum dan Lynn Streeter. Dalam konteks aplikasinya ke pencarian informasi, metode LSA ini juga disebut sebagai *Laten Semantic Indexing* (LSI).

LSA dapat ditafsirkan sebagai cara yang cepat dan praktis untuk mendapatkan perkiraan perkiraan *substitutability* kontekstual penggunaan kata-kata dalam segmen teks yang besar yang belum ditentukan makna kesamaan antara kata-kata dan segmen teks yang mungkin mencerminkan suatu hubungan tertentu. Sebagai metode praktis untuk mengkarakterisasi arti dari kata, LSA menghasilkan ukuran hubungan kata-kata, bagian kata dan bagian-bagian yang berkorelasi dengan beberapa fenomena kognitif manusia yang melibatkan asosiasi atau kesamaan semantik. Konsekuensi praktis dari metode LSA ini, memungkinkan kita untuk sangat mendekati penilaian manusia untuk menilai kesamaan makna antara kata dan secara objektif memprediksi konsekuensi dari keseluruhan kata berdasarkan kesamaan antara bagian-bagian kata serta perkiraan yang kata yang sering muncul. Sedangkan *Probabilistic Latent Semantic Analysis* (PLSA) adalah sebuah algoritma yang diterapkan untuk

memperkirakan makna sekumpulan teks menjadi suatu *cluster* atau kelompok (kategori) tertentu sehingga mempermudah para analis untuk menarik suatu kesimpulan dari pengelompokkan yang terbentuk. Secara umum metode PLSA menggabungkan teori klasik tentang *vector space model*, *Singular Value Decomp*osition (SVD) serta model variabel latent, yang diformulasikan kedalam suatu bentuk model peluang dengan tujuan untuk mendapatkan suatu kelompok (*latent*) dari sekumpulan teks (*bag of words*). Aplikasi PLSA ini dapat diterapkan dalam analisis *sentiment* pada pasar saham, analisis kemiripan dokumen untuk mendeteksi plagiarism, analisis trending topik pada media social. Permasalahan yang timbul dalam penggunaan metode LSA ini adalah adanya faktor polysemi dalam pengelompokkan kata (Hofmann, 2001). Permasalahan polysemi pada kata dapat diatasi dengan mengunakan varian dari LSA yang dikenal sebagai *Probabilistic Latent Semantic Analysis* (PLSA).

Latent class Metode PLSA pada dasarnya merupakan model campuran dari model *latent class* dengan kata lain model *latent class* untuk data teks, PLSA sendiri merupakan salah satu model *based clustering* yang bertujuan untuk membentuk *cluster* berdasarkan model peluang statistik, berbeda dengan metode *cluster* yang konvensional metode ini dapat dievaluasi berdasarkan ukuran statistik tertentu. Keutamaan metode PLSA sendiri dapat mereduksi dimensi matriks *term* yang terbentuk dari sekekumpulan teks yang direpresentasikan dalam sebuah variabel latent, sehingga ukuran dimensi matriks kemunculan *term* pada metode LSA dapat direduksi (Hofmann, 1999). Baik model LSA dan PLSA mengabaikan urutan kata (word ordering) dalam proses analisisnya, hal ini menjadi masalah dikarenakan beberapa term tertentu akan memiliki makna yang jauh berbeda jika diperhatikan urutannya sehingga memungkinkan pengelompokkan term kedalam suaru topik menjadi tidak tepat, LDA menjadi salah satu solusi untuk mengatasi pengelompokaan term majadi topik tertentu dengan memperhatikan urutan kata pada proses pembentukkanya melalui mekanisme model campuran (*mixture*).

METODE PENELITIAN

Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif dari sekumpulan *corpus*, Ide dasarnya adalahbahwa dokumen dapat direpresentasikan sebagai model campuran dari berbagai topik yang disebut juga laten, di mana setiap topik dikarakateristikan oleh kata. LDA mengasumsikan proses generatif berikut untuk setiap dokumen $\boldsymbol{w}$ dalam sebuah corpus$\boldsymbol{D}$ adalah sbb:

1. Pilih $N \sim Pissson(\xi)$,
2. Pilih $\boldsymbol{\theta} \sim Dir(\alpha)$,
3. Untuk setiap N kata w_n,

a. Pilih Topik $z_n \sim Multinomial(\boldsymbol{\theta})$,

b. Pilih sebuah kata w_n dari $p(w_n \mid z_n, \beta)$.

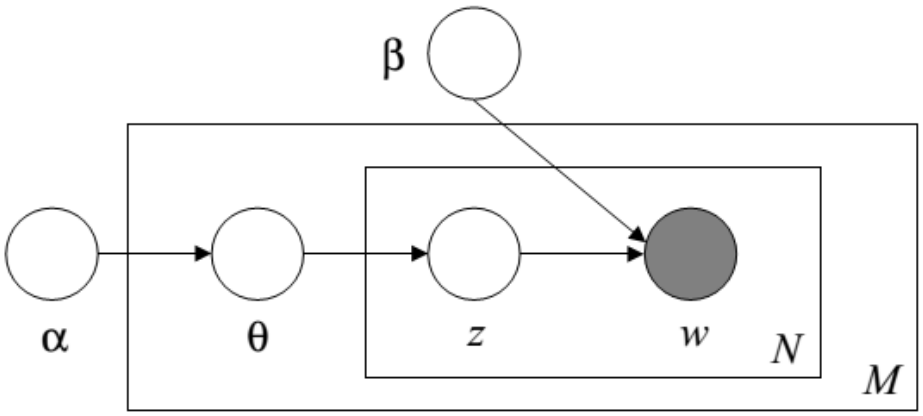
Beberapa asumsi penyederhanaan yang dibuat dalam model dasar LDA, Pertama, distribusi dari topik (latent) diketahui mengikuti k distribusi Dirichlet. Kedua, probabilitas kata adalah matriks $\boldsymbol{\beta}$ berukuran $k \times V$ yang mana $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$. Sedangkan k distribusi Dirichlet memiliki fungsi densitas sbb:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \quad (1)$$

Bentuk distribusi bersama dari Topik mixture θ dari *N* topik ***z*** dan *N* kata ***w*** besyarat α dan β adalah:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \quad (2)$$

Representasi model LDA jika digambarkan dalam sebuah diagram dapat digambarkan sbb:



Gambar 1. Representasi Model LDA.

Bentuk distribusi marginal $p(\mathbf{w} \mid \alpha, \beta)$ didapat dengan menintegrasikan persamaan (2) terhadap θ sbb:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \quad (3)$$

Akhirnya , perkalian densitas marginal untuk dari sebuah dokumenakan memperoleh probabilitas marginal sebuah corpus sbb:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad (4)$$

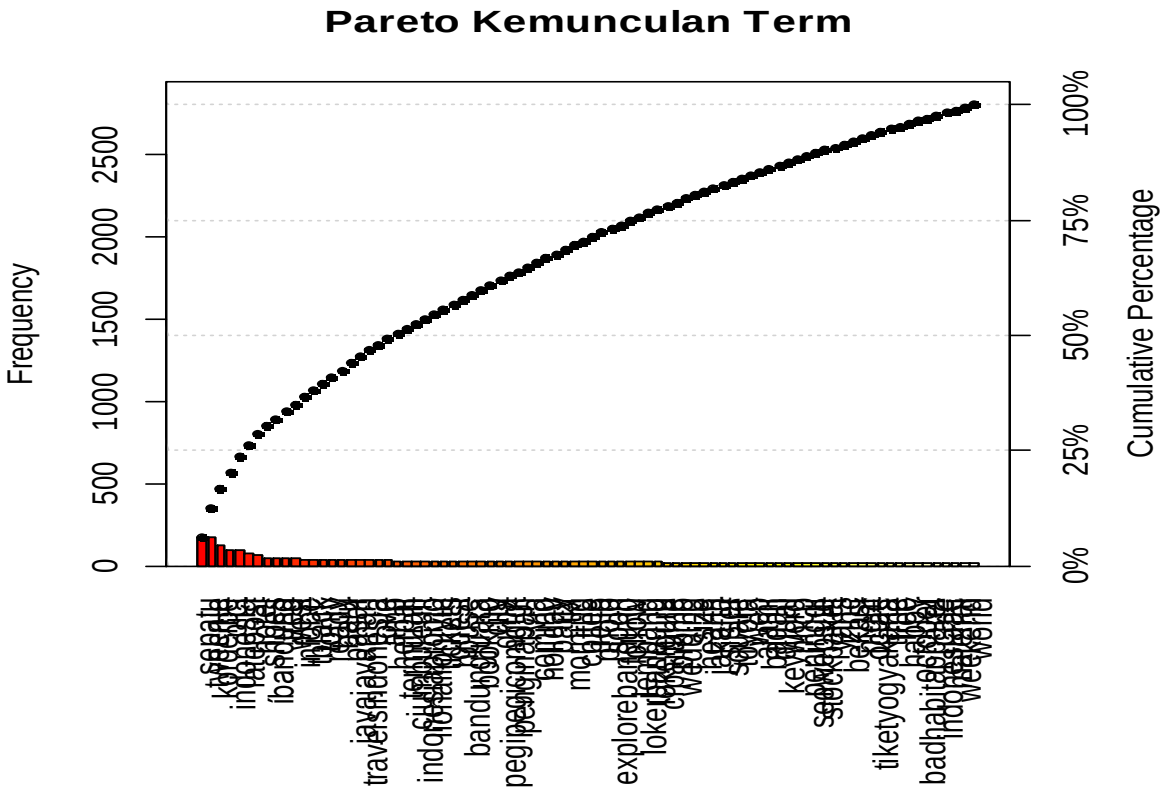
Model LDA dapat direpresentasikan sebagai model grafis probabilistik seperti pada Gambar 1. Gambar 1 menjelaskan, ada tiga tingkat sebagai representasi LDA. Parameter parameters α dan β sebagai corpuslevel parameter, diasumsikan disampel sekali dalam proses menghasilkan sebuah *corpus*. Variabel θ_d adalah variabel pada tingkat dokumen, disampel sekali per dokumen. Akhirnya, variabel z_{dn} dan w_{dn} yang merupakan variabel pada tingkat

yang disampel sekali untuk setiap kata dalam setiap dokumen. Struktur model pada Gambar 1 merupakan model statistik Bayesian yang disebut juga model *Bayesian hierarchical* (Gelman et al., 1995), atau lebih tepatnya model *asconditionally independent hierarchical models* (Kass dan Steffey, 1989) dan *asparametric empirical Bayes models* (Morris, 1983).

HASIL DAN PEMBAHASAN

Pada penelitian ini akan dipergunakan data dari twitter dengan mengambil sampel sebanyak 1500 tweet dengan kata kunci #Bandung, Pada proses awal dilakukan tahapan pembersihan data teks dengan cara melakukan pembuangan kata yang tidak penting melalui tahapan *stopword*, dilanjutkan dengan proses *steamming*. Setelah tahapan *cleaning* data text sudah dilakukan langkah selanjutnya adalah dengan membuat matriks kemunculan kata berdasarkan tweet yang terjadi, proses ini dilakukan dengan bantuan software R.

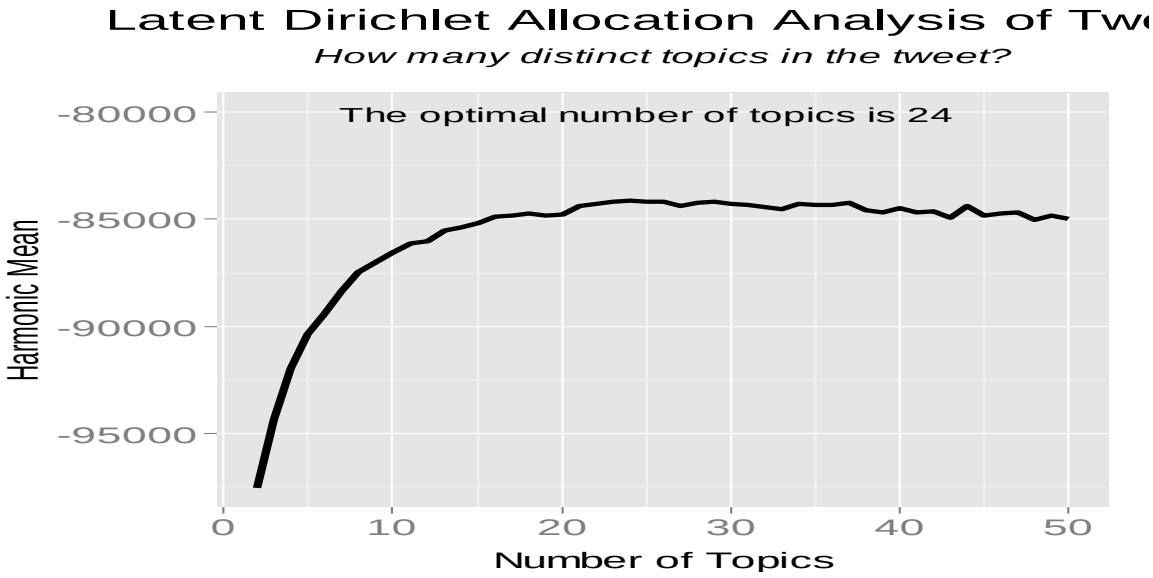
Hasil analisis awal menunjukkan *term* data memiliki tingkat kejadian yang paling tinggi jika dibandingan dengan *term* kata lainnya hal ini dapat dilihat pada Gambar 2. Kemunculan *term-term* yang sering muncul juga dapat dilihat dari gambar *wordcloud* pada Gambar 3.



Gambar 2. Pareto*Terms*

LDA akan dipergunakan pada pengidentifikasian Topiks pada tweet #RDatamining. Adapun langkah awal dari proses LDA adalah menetapkan jumlah topik awal yang selanjutnya akan diproses lebih lanjut dengan Algoritma LDA. Jumlah topik ditentukan berdasarkan plot antara banyaknya topik dengan nilai *likelihood* dari model LDA (Gambar 4)

Gambar 3. *Wordcloud Terms.*

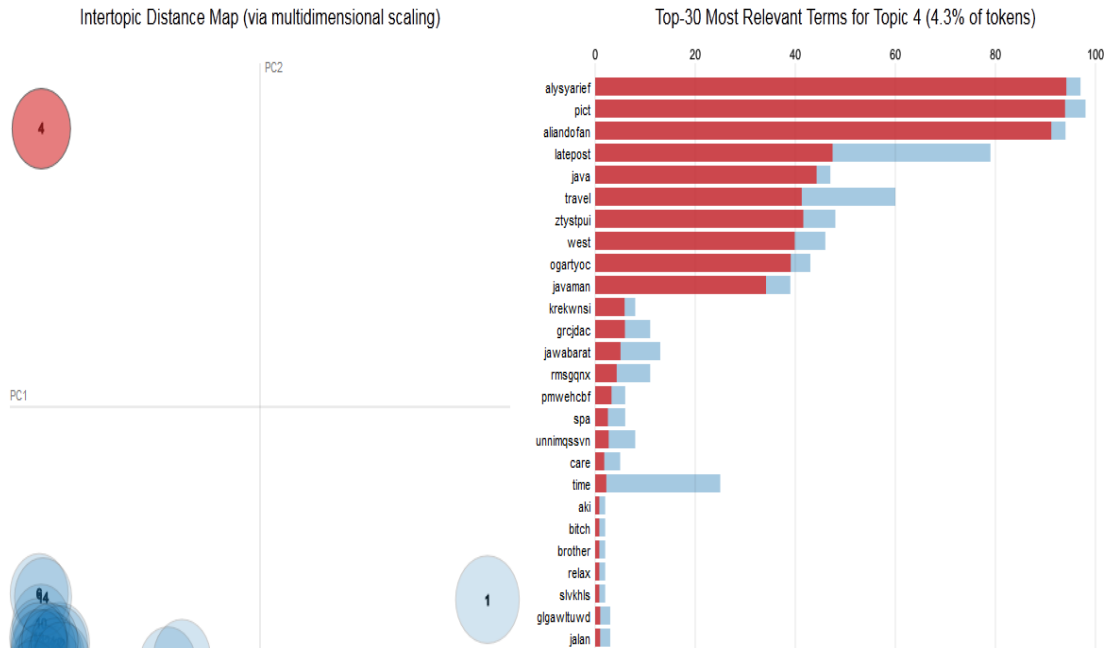


Gambar 4. Plot jumlah topik terhadap *Likelihood.*

Bentuk visualisai dari masing-masing topik dan kata dapat dilihat dalam diagram yang tertera pada Gambar 5.

KESIMPULAN

Metode pengklasteran LDA mengelompokkan data twitter dengan kata kunci bandung menghasilkan 24 buah Topik, Metode LDA yang diimplementasikan pada R masih terbatas dalam hal jumlah dimensi dari *term* yang terbentuk serta jumlah term yang dapat diretrieve dari server twitter, Keterbatasan lainnya bahwa metode yang ada saat ini hanya berlaku untuk bahasa inggris, untuk dapat diterapkan pada bahasa Indonesia diperlukan algoritma dan database kata dasar dalam bahasa Indonesia. Sehingga hal ini menjadi saran bagi peneliti lainnya untuk dapat mengimplementasikan metode ini dalam bahasa Indonesia.



Gambar 5. Plot Visualisai LDA

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Departemen Statistika FMIPA UNPAD yang telah memberi dukungan financial terhadap penelitian/makalah ini.

REFERENSI

[1] Anglin, J. M. (1970). *The Growth of Word Meaning*. Cambridge, MA.: MIT Press.
[2] Anglin, J. M., Alexander, T. M., & Johnson, C. J. (1996). Word learning and the growth of potentially knowable vocabulary. Submitted for publication.
[3] Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. In D. Harman (Ed.), The Second Text Retrieval Conference (TREC2), *National Institute of Standards and Technology Special Publication* 500-215, pp. 105-116.
[4] Dumais, S. T. & Nielsen, J. (1992). *Automating the assignment of submitted manuscripts to reviewers*. In N. Belkin, P. Ingwesen, & A. M. Pejtersen (Eds.)
[5] Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. York, Association for Computing Machinery.
[6] Hutomo, A. dan Zulhanif (2013). *Analisis Keluhan Penumpang PT. Kereta Api Indonesia (Persero) Menggunakan LSA dan Analisis Korespondensi*. Univesitas Padjadjaran.
[7] Hofmann. T. (1999). Probabilistic Latent Semantic Indexing, *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval* (SIGIR-99).
[8] Hofmann, T., Puzicha, J., dan Jordan, M. I. (1999). Unsupervised learning from dyadic data. *In Advances in Neural Information Processing Systems*, Vol. 11, MIT Press.
[9] Saul, L. & Pereira, F. (1997). Aggregate and mixed-order Markov models for statistical language processing. *Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing*, pp. 81-89.