UNIVERSITAS INDONESIA

EKSTRAKSI TOPIK UTAMA HARIAN DARI PORTAL BERITA INDONESIA *ONLINE* MENGGUNAKAN *SINGULAR VALUE DECOMPOSITION*

SKRIPSI

ASHARI NURHIDAYAT
0706261562

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
PROGRAM SARJANA MATEMATIKA
DEPOK
JUNI 2012

UNIVERSITAS INDONESIA

EKSTRAKSI TOPIK UTAMA HARIAN DARI PORTAL BERITA INDONESIA *ONLINE* MENGGUNAKAN *SINGULAR VALUE DECOMPOSITION*

SKRIPSI

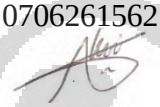
Diajukan sebagai salah satu syarat untuk memperoleh gelar sarjana sains

ASHARI NURHIDAYAT
0706261562

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
PROGRAM SARJANA MATEMATIKA
DEPOK
JUNI 2012

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama	: Ashari Nurhidayat
NPM	: 0706261562
Tanda Tangan	: 
Tanggal	: 15 Juni 2012

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :
nama : Ashari Nurhidayat
NPM : 0706261562
program studi : Sarjana Matematika
judul skripsi : Ekstraksi Topik Utama Harian dari Portal Berita Indonesia *Online* Menggunakan *Singular Value Decomposition*

telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Sarjana Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Indonesia

DEWAN PENGUJI

Pembimbing	: Dr. rer. nat Hendri Murfi M. Kom	( )
Penguji	: Dr. Al Haji Akbar B, M.Sc	( )
Penguji	: Dr. Kiki Ariyanti S, M.Si	( )
Penguji	: Dra. Siti Aminah M.Kom	( )

Ditetapkan di : Depok
Tanggal : 15 Juni 2012

KATA PENGANTAR

Alhamdulillah, tiada kalimat yang lebih indah sebagai rasa syukur penulis atas segala nikmat, rahmat dan karunia-Nya. Shalawat dan salam teruntuk suri tauladan yang sempurna, Nabi Muhammad SAW. Pada saat ini penulis telah dapat menyelesaikan tugas akhir ini. Penulis sadar, bahwa tugas akhir ini tidak lepas dari bantuan dan dukungan dari berbagai pihak. Pada kesempatan ini, penulis menghaturkan terima kasih sebesar-besarnya kepada:

(1) Dr. rer. nat Hendri Murfi M.Kom selaku pembimbing yang telah memberikan segala bimbingan dan masukan-masukan sehingga tugas akhir ini dapat tersusun.
(2) Dra. Yahma Wisnani M.Kom, selaku pembimbing akademik.
(3) Prof. Dr. Djati Kerami, Dr. Yudi Satria, Drs. Suryadi M.T M.T, Dr. Al Haji Akbar M.Sc, Dr. Kiki Ariyanti S, M.Si, Dr. Sri Mardiyati M.Kom, Dra. Siti Aminah M.Kom yang telah hadir dalam SIG1 dan SIG2 serta menguji dalam kolokium untuk memberikan perbaikan atas tugas akhir ini.
(4) Seluruh staf pengajar departemen Matematika UI atas ilmu, pengalaman dan bimbingan yang diberikan.
(5) Seluruh karyawan departemen Matematika UI atas bantuannya mempermudah proses perkuliahan.
(6) Ibu dan bapak yang tiada henti mendoakan dan memberikan kasih sayang serta dukungan yang tiada terputus.
(7) Kakak dan adik-adik yang selalu memberikan dukungan dan semangat.
(8) Ibu Istiqomah beserta anak-anaknya atas bantuan dan nasihat yang diberikan.
(9) Nasrul Latif S.Psi, Dwi Wahyu Prabowo S.Si, Salman El Farisi S.Kom yang telah memberikan dukungan dan nasehat selama menempuh kegiatan kampus
(10) Eka Mustikawati, S.S atas pembuatan topik utama secara manual dan bantuan koreksi tugas akhir ini.
(11) Bang Susanto atas bantuan *keyboard*, *mouse* dan pembelajaran Linux yang diberikan.

(12) Keluarga besar HMD Matematika UI 2009 dan BEM FMIPA UI 2010, terutama untuk CT-BPH yang telah mewarnai kehidupan penulis di dunia mahasiswa

(13) Keluarga BEM UI 2011, untuk PI dan BPH, terutama tim Bidang Keilmuan yang tidak bisa disebutkan satu-satu.

(14) Saudara seperjuangan "KIAM","Bintang Kecil", "Fathan mubina", "Brownies", dan "50$p0L1c10u5 2011" atas perjuangan dan ikatan hati yang memberikan semangat baru saat kejenuhan menghinggap.

(15) Kawan-kawan dan saudara seperjuangan di matematika angkatan 2007, baik yang tetap maupun yang 'memilih' jalan lain.

(16) Rekan-rekan asisten laboratorium Matematika UI.

(17) Rekan-rekan Matematika angkatan 2004,2005,2006 atas pengalaman dan pembelajarannya serta angkatan 2008, 2009,2010 dan 2011.

(18) Seluruh mahasiswa departemen matematika yang bersama menyusun tugas akhir, baik skripsi maupun thesis.

(19) Kelompok JKMI 18 dan 16 beserta Dr. Lukman dan pak Anton yang membimbing penulis selama setahun terakhir.

(20) *Ranggers* DKI yang tiada lelah memberikan yang terbaik untuk perbaikan ummat.

(21) The Micins dan adik-adik Vidatra 34 yang telah memberikan semangat luar biasa. Semoga kalian bisa menggapai cita-cita dan mimpi.

(22) Saudara dalam kontrakan matematika (koma), Dimas, Arif, Ridwan, Ali, Sigap, Bayu, Kemal, Marcel.

penulis juga mengucapkan terima kasih kepada pihak-pihak yang tidak mungkin dituliskan satu persatu atas bantuan dalam menyusun tugas akhir ini dan memberikan warna dalam kehidupan penulis dalam dunia mahasiswa. Akhir kata, penulis memohon maaf jika ada kesalahan dan kekurangan dalam tugas akhir ini. Penulis berharap tugas akhir ini dapat bermanfaat bagi perkembangan ilmu pengetahuan.

Penulis

2012

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI

TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama	: Ashari Nurhidayat
NPM	: 0706261562
Program Studi	: Sarjana Matematika
Departemen	: Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis karya	: Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul :

EKSTRAKSI TOPIK UTAMA HARIAN DARI PORTAL BERITA INDONESIA *ONLINE* MENGGUNAKAN *SINGULAR VALUE DECOMPOSITION*

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalih media/ format-kan, mengelola dalam bentuk pangkalan data (database), merawat dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : 15 Juni 2012
Yang menyatakan

(Ashari Nurhidayat)

ABSTRAK

Nama : Ashari Nurhidayat
Program Studi : Sarjana Matematika
Judul : Ekstraksi Topik Utama Harian dari Portal Berita Indonesia *Online* Menggunakan *Singular Value Decompositon*

Ekstraksi topik adalah kegiatan untuk mendapatkan topik dalam kumpulan dokumen berita. Ekstraksi topik memiliki peran yang penting untuk mendapatkan maksud dari keseluruhan dokumen teks tersebut. Metode yang umum digunakan dalam *machine learning* untuk pencarian topik utama adalah *unsupervised learning*, dimana topik diekstraksi dari kumpulan dokumen tanpa bergantung pada label dokumen. Salah satu metode yang dapat digunakan untuk mengekstraksi topik dari kumpulan dokumen berita yaitu *latent semantic analysis* (LSA). LSA mengaplikasikan teknik *singular value decomposition* (SVD) untuk mendapatkan hubungan kata dengan topik dalam kumpulan dokumen berita. Pada skripsi ini, dibahas mengenai implementasi metode LSA pada kumpulan dokumen dari portal berita *online* berbahasa Indonesia. Selanjutnya, keluaran metode LSA dibandingkan dengan hasil ekstraksi topik secara manual untuk menunjukkan keberhasilan metode LSA.

Kata Kunci : Ekstraksi Topik, LSA, SVD, *Machine Learning*, *Unsupervised Learning*.

xiv+52 halaman : 15 tabel, 14 gambar, 9 lampiran
Daftar Pustaka : 11 (1990-2011)

ABSTRACT

Name : Ashari Nurhidayat
Program Study : Mathematics
Title : Daily Main Topic Extraction from Indonesia News Online Portals using Singular Value Decomposition

Topic extraction is an activity to get a topic from text document collection. Topic extraction is very important in order to find out the meaning of those whole text document. The general method used in machine learning for finding the main topic is unsupervised learning, where a topic is extracted from the document collection without depending on document labels. One of Methods which can be used for extracting a topic from text document collection is latent semantic analysis (LSA). Furthermore, LSA using LSA to show a relation between words and topic in their organizer document collection. In this skripsi, the implementation of LSA method in documents collection from Indonesian online news portal discussed. Furthermore, LSA method output compared with manual extraction to demonstrate the success of LSA.

Key Words : Topic Extraction , LSA, SVD, Machine Learning, Unsupervised Learning.

xiv+52 pages : 15 tables, 14 pictures, 9 attachments
Bibliography : 11 (1990-2011)

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan dunia teknologi informasi begitu cepat ikut mendorong peningkatan pengguna internet. Berdasarkan data yang dilansir Bank Dunia, pengguna internet di Indonesia tumbuh dari 0,93% ditahun 2000 menjadi 9,1% ditahun 2010[1]. Hasil survey oleh Nielsen terhadap pengguna teknologi di Asia Tenggara menyebutkan 68% pengguna internet di Indonesia mengakses internet melalui perangkat bergerak[2]. Selanjutnya, survey yang dilakukan Yahoo! menyebutkan sebanyak 61% pengguna internet mengakses portal berita *online*[3].

Seperti yang dilansir detik.com, kecenderungan masyarakat untuk mengakses portal berita *online* semakin besar[4]. Dengan mengakses berita *online*, masyarakat medapatkan berita teraktual dimana saja dan kapan saja tanpa harus menunggu proses pencetakan. Hal ini memberikan tantangan terhadap media-media konvensional atau media cetak. Sehingga kantor-kantor berita berlomba-lomba untuk dapat menyajikan informasi teraktual di dalam portal beritanya.

Upaya untuk memberikan berita teraktual menjadikan arus informasi menjadi sangat cepat. Sebagai contoh, berdasarkan pengamatan penulis, detik.com mampu untuk mewartakan hingga 200 berita dalam satu hari[5]. Selain itu, berkembangnya jumlah portal berita *online* di Indonesia ikut mempersulit masyarakat untuk dapat mendapatkan informasi secara menyeluruh.

Informasi yang utuh dapat membantu pembaca atau masyarakat mengetahui topik utama dari pemberitaan media dalam jangka waktu tertentu. Topik utama dari kumpulan berita ini menjadi sangat penting untuk mengetahui *trend* yang terjadi dalam waktu tertentu. Selain itu, topik utama ini juga dapat dijadikan dasar untuk mengambil kebijakan dimasa depan.

---

[1] Dipublikasikan di http://www.google.com/publicdata/explore diakses pada 27 Maret 2012 pukul 12.00 WIB

[2] *Nielsen Southeast Asia Digital Consumers Report 2011*, dipublikasikan di http://www.thejakartapost.com/ diakses pada 27 Maret 2012 pukul 12.00 WIB

[3] Yahoo survey for internet access 2011

[4] Dipublikasikan di http://www.detik.com/ diakses pada 27 Maret 2012 pukul 12.00 WIB

[5] Ibid.

Metode untuk mencari topik utama secara manual dilakukan dengan membaca seluruh artikel. Selanjunya dengan menentukan seluruh kalimat topik dalam dokumen, pembaca dapat menentukan topik utamanya. Metode ini membutuhkan sumber daya yang besar dan kemampuan intuisi yang baik dari pembaca. Oleh karena itu dibutuhkan metode lain yang dapat dijalankan secara otomatis untuk mengekstraksi topik utama dari seluruh dokumen berita.

Deerwester, Dumais, Furnas, Launder, Harshman (1990) memperkenalkan teknik *Latent Semantic Analysis* (LSA) untuk mengekstraksi topik utama dari kumpulan dokumen berita. LSA merepresentasikan kumpulan dokumen menjadi matriks kata-dokumen. Setiap anggota dari matriks kata-dokumen menunjukkan frekuensi kata dalam dokumen. Selanjutnya digunakan *Singular Value Decomposition* (SVD) untuk mengurangi dimensi matriks kata-dokumen.

Menurut Golub (1996), SVD mendekomposisi matriks berukuran mxn menjadi dua matriks ortogonal berukuran mxm dan nxn serta satu matriks berukuran mxn yang anggotanya bernilai 0 kecuali diagonal utamanya bernilai *singular* dari matriks kata-dokumen. SVD dapat di implementasikan terhadap sembarang matriks.

Dengan asumsi bahwa setiap kumpulan dokumen memiliki struktur semantik tersembunyi berupa topik, maka dengan SVD pada matriks kata-dokumen, hubungan semantik tersembunyi atau topik dapat diekstraksi (Murfi, 2010)

1.2 Perumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah yang diangkat dalam penelitian tugas akhir ini adalah bagaimana mengimplementasikan LSA dengan SVD untuk mengekstraksi topik-topik utama harian dari kumpulan dokumen berita *online* berbahasa Indonesia.

1.3 Tujuan Penelitian

Penelitian ini bertujuan mengimplementasikan metode LSA dengan SVD untuk mengekstraksi topik-topik utama harian dari kumpulan dokumen berita *online* berbahasa Indonesia.

1.4 Metodologi Penelitian

Penelitian pada tugas akhir ini melalui langkah-langkah berikut :

(a) Perumusan Masalah dan Studi Literatur

Pada tahap awal penelitian, dirumuskan masalah penelitian beserta tahap-tahap penyelesaiannya. Studi literatur dilakukan untuk mendapatkan teori-teori yang mendukung topik penelitian yang berasal dari jurnal, buku dan media lainnya.

(b) Pengumpulan Data

Pada tahap ini, akan dilakukan penentuan sumber data yang akan digunakan dengan bantuan file berformat XML berupa RSS dari beberapa portal berita online berbahasa Indonesia. Portal berita yang akan digunakan adalah :

a. Kompas.com[6]
b. Okezone.com[7]
c. Detik.com[8]
d. Vivanews.com[9]
e. Antaranews[10]
f. Republika online[11]
g. Media Indonesia[12]

(c) Implementasi Algoritma dan Simulasi

Selanjutnya, data yang diperoleh setiap hari diimplementasikan pada metode LSA denngan perangkat lunak berbahasa Python. Perangkat lunak ini digunakan untuk membangun matriks kata-dokumen dari data yang telah didapat, selanjutnya dengan mengaplikasikan metode LSA untuk mencari topik utama dalam kumpulan dokumen berita nasional selama beberapa hari.

[6] http://www.kompas.com/
[7] http://www.okezone.com/
[8] http://www.detik.com/
[9] http://www.vivanews.com/
[10] http://www.antaranews.com/
[11] http://www.republika.co.id/
[12] http://www.mediaindonesia.com/

(d) Interpretasi dan Analisa Hasil

Tahap akhir adalah dengan membangun interpretasi topik utama dari keluaran meode LSA. Topik-topik utama tersebut kemudian dianalisa dan dibandingkan dengan hasil ekstraksi menggunakan metode manual.

1.5 Batasan dan Ruang Lingkup Masalah

Penelitian ini dibatasi dengan asumsi sebagai berikut :

(a) Dokumen berita memiliki struktur semantik tersembunyi yang disebut sebagai topik;
(a) Topik dapat disimpulkan dari hubungan antara kata dengan dokumen;
(b) Kata memiliki kaitan dengan topik;
(c) Interpretasi kalimat topik dilakukan secara manual menggunakan kata-kata keluaran metode LSA;
(d) Dokumen berita berasal dari portal berita berbahasa Indonesia yang disebutkan pada subbab 1.4 poin (b);
(e) Dokumen berita yang digunakan dalam simulasi didapat dari satu kali pengambilan setiap harinya selama bulan Mei.

BAB 2
LANDASAN TEORI

Pada bab ini akan dibahas tentang teori-teori yang mendukung penggunaan metode LSA. Pembahasan diawali dengan *machine learing,* ekstraksi topik utama dan dilanjutkan dengan *singular value decomposition* (SVD) serta teori matematis yang mendasari SVD.

2.1 *Machine Learning*

Machine learning merupakan bagian dari *artificial intellegence* (AI) atau kecerdasan buatan yang berfokus pada algoritma komputer (mesin) untuk 'belajar' dari sejumlah data yang disebut *data training.* Algoritma tersebut digunakan untuk menyimpulkan informasi tentang sifat-sifat dan pola data (Segaran, 2007). Selanjutnya, informasi tersebut dapat digunakan untuk memprediksi data lain pada masa selanjutnya. Hal ini dimungkinkan karena hampir semua data yang tertentu memiliki pola, sehingga memungkinkan mesin untuk menggeneralisasi pola tersebut.

Menurut Ghahramani (2004), *machine learning* dapat dibagi kedalam empat kelompok berdasarkan perlakuan terhadap data yang diberikan.

(a) *Supervised Learning*

Pada teknik *supervised learning,* barisan *data training* yang diberikan memiliki barisan harapan keluaran atau label. Harapan keluaran ini dapat berupa kelas label atau bilangan riil. Beberapa metode untuk mendapatkan label diantaranya *neural network, desicion trees, support-vector machine* dan *Bayesian filtering.* Tujuan teknik ini adalah menjadikan mesin dapat belajar untuk memberikan keluaran yang terbaik untuk *data training* baru berdasarkan label.

(b) *Reinforcement Learning*

Dalam *reinforcement learning,* mesin berinteraksi dengan lingkungan yang memberikan barisan aksi. Aksi ini dipengaruhi kondisi awal lingkungan dan memberikan hasil kepada mesin berupa barisan skalar *reward* atau *punishment.* Tujuan metode ini menjadikan mesin dapat 'belajar' untuk

mengoptimalkan *reward* dan meminimalkan *punishment* selama mesin tersebut bekerja. *Reinforcement learning* berdekatan dengan bidang teori keputusan (*desicion theory*) dan teori kontrol (*control theory*).

(c) *Generalized Reinforcement Learning*

Jenis ketiga ini erat kaitannya dengan *game theory* dan merupakan generalisasi dari teknik *reinforcement learning.* Seperti metode *reinforce learning,* mesin membutuhkan masukan, menghasilkan aksi dan mendapatkan *reward* atau *punishment.* Akan tetapi lingkungan yang berinteraksi dengan mesin bukanlah 'dunia statis', melainkan juga mesin lain yang dapat merasa, bertindak, mendapat *reward* atau *punishment* dan 'belajar'. Tujuan teknik ini adalah agar mesin dapat memaksimalkan *reward* dan meminimalkan *punishment* pada setiap aksi yang dilakukan saat ini hingga masa depan.

(d) *Unsupervised Learning*

Berbeda dengan teknik lainnya, *unsupervised learning* tidak membutuhkan label untuk 'mempelajari' dan membuat prediksi dari kumpulan *training data* serta tidak melakukan interaksi dengan lingkungan. Teknik ini bertujuan untuk mendapatkan representasi dari *data training* yang dapat digunakan untuk memberikan suatu keputusan, memprediksi masukan selanjutnya dan lainnya. Teknik ini dapat digunakan untuk menemukan pola struktur yang bebas dari *noise.* Metode yang menggunakan teknik *unsupervised learning* diantaranya *non-negative matrix factorization, self-organizing map* dan *latent semantic analysis.*

2.2 Ekstraksi Topik Utama

Topik dalam kamus besar bahasa Indonesia (KBBI, 2008) merupakan kata yang memiliki makna sebagai pokok pembicaraan dalam diskusi, ceramah, karangan, dan sebagainya. Topik juga dapat diartikan sebagai bahan diskusi atau hal yang menarik perhatian umum pada waktu akhir-akhir ini.

Dengan menggunakan asumsi pada subbab 1.5 poin (a), (b) dan (c), maka dengan menggunakan metode tertentu, kita dapat mengekstraksi informasi berupa topik utama pada kumpulan dokumen. Metode tradisional yang dilakukan untuk

mengekstraksi topik utama dari kumpulan dokumen adalah dengan cara manual. Seperti penjelasan dalam subbab 1.1, cara ekstraksi manual dilakukan dengan membaca seluruh dokumen, selanjutnya dengan menentukan seluruh kalimat topik dan menarik kesimpulan maka topik utama dapat diekstraksi.

Honkela (2004) menyatakan pencarian topik telah digunakan secara luas dengan menggunakan analisa statistika dan matematika dari kumpulan bahasa natural. Permasalahan utamanya adalah bagaimana mengubah bahasa tulisan agar dapat diproses oleh komputer. Mengubah masukan berupa simbol-simbol bahasa berupa kata-kata untuk diolah dalam algoritma numerik. Kemiripan dari suatu kata belum tentu memiliki hubungan yang dekat. Sebagai contoh, kata 'kaca', 'kacau' dan 'gelas'. Kata 'kaca' dan 'kacau' memiliki kedekatan cara penyebutan (fonetik) sehingga memiliki kedekatan representasi numerik. Sedangkan, kata 'kaca' dan 'gelas' memiliki kedekatan makna namun tidak memiliki kedekatan dalam representasi numerik.

Salah satu representasi numerik yang digunakan untuk mendapatkan hubungan semantik adalah dengan memperhatikan konteks kalimat dimana kata tersebut berada. Beberapa metode yang umum digunakan untuk mendapatkan hubungan semantik ini diantaranya :

(a) *Latent Semantic Analysis*

Metode ini merepresentasikan kumpulan dokumen dalam bentuk matriks kata-dokumen yang selanjutnya dengan SVD didapatkan hubungan kata dengan topik. Metode ini akan dibahas lebih lanjut dalam bab 3 dan bab 4.

(b) *Self-Organizing Map of Words*

Self-organized map telah digunakan untuk mengekstrak hubungan semantik kata dalam *artificial short sentence* dan *Grimm Fairy Tales.* Analisa kata dengan *self-organized map* menggunakan peta satu dimensi yang bertujuan untuk mencari kata-kata yang memiliki hubungan sinonim. Hasilnya disebut peta kategori kata.

(c) *Independent Componet Analysis* (ICA)

ICA termasuk dalam teknik *unsupervised learning*. ICA memberikan fitur-fitur berbeda yang merefleksikan pembentukan dan kategori semantik. ICA menggunakan model persamaan :

$$\boldsymbol{x} = A\boldsymbol{s} \qquad (2.1)$$

dimana $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^T$ merupakan vektor dari variabel acak yang diobservasi. Kemudian, $\boldsymbol{s} = (s_1, s_2, \dots, s_n)^T$ merupakan vektor dari variabel *latent independent* atau *independent component.* Matriks *A* merupakan suatu matriks konstan yang disebut *mixing matrix.* Dalam metode ICA, dengan memiliki matriks $\boldsymbol{x}$ yang merupakan data observasi atau *data training*, akan diestimasi matriks *A* dan vektor $\boldsymbol{s}$ yang menunjukkan hubungan semantik.

2.3 Faktorisasi Matriks

Faktorisasi matriks merupakan proses pemecahan atau penguraian suatu matriks menjadi beberapa matriks. Matriks-matriks hasil faktorisasi biasanya memiliki struktur tertentu dimana membuat beberapa operasi akan menjadi lebih sederhana (efisien dari segi komputasi), dan atau jumlah komponen yang lebih sedikit (efisien dari segi memori). Beberapa contoh matriks hasil faktorisasi adalah matriks triangular (segitiga atas, segitiga bawah, diagonal), matriks ortogonal atau matriks yang memiliki *rank* yang lebih kecil.

Secara umum, metode faktorisasi matriks dibagi menjadi dua kelompok, yaitu *direct method* dan *approximation method*. *Direct method* merupakan teknik yang secara teori memberikan nilai eksak dengan jumlah langkah terbatas. Contoh faktorisasi matriks dengan *direct method* adalah faktorisasi LU, faktorisasi *Cholesky* dan faktorisasi QR. Sedangkan *approximation method* menggunakan suatu perkiraan solusi awal dan dilanjutkan dengan iterasi yang memberikan solusi hasil lebih baik. Tujuan metode ini untuk mendapatkan cara meminimalkan perbedaan antara solusi perkiraan (*approximation*) dan solusi eksak. Contoh faktorisasi matriks dengan *approximation method* adalah *single value decomposition* (SVD), *matrtix factorization* (MF) dan *non-negative matrix factorization* (MNF).

2.4 *Norm, Nullity* dan *Rank* Matriks

Sebelum dijelaskan tiga sifat dari matriks, yaitu *norm, nullity* dan *rank,* akan didefinisikan vektor bebas linear dan ruang hasil kali dalam.

Definisi 2.1

Misalkan $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_k}\}$ merupakan himpunan vektor-vektor di $\mathbb{R}^k$. Himpunan tersebut dikatakan **bebas secara linear** (***linearly independent***) jika berlaku:

$$\mathbf{0} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k} \qquad (2.2)$$

maka $\alpha_i = 0$, untuk setiap $i = 1,2, \dots, k$. Selain itu, himpunan vektor tersebut dikatakan **tidak bebas secara linear** atau **bergantung secara linear** (***linearly dependent***) (Burden, 2011, hal: 564).

Teorema berikut memberikan cara untuk menyatakan vektor yang tidak bebas secara linear atau suatu vektor merupakan kombinasi linear dari himpunan vektor yang bebas linear.

Teorema 2.2

Misalkan $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_n}\}$ merupakan himpunan dari n vektor yang bebas secara linear di $\mathbb{R}^n$, maka untuk setiap vektor di $\boldsymbol{x} \in \mathbb{R}^n$ terdapat kumpulan konstanta yang unik $\beta_1, \beta_2, \dots, \beta_n$ memenuhi :

$$\boldsymbol{x} = \beta_1 \boldsymbol{v_1} + \beta_2 \boldsymbol{v_2} + \cdots + \beta_n \boldsymbol{v_n} \qquad (2.3)$$

(Burden, 2011, hal: 564).

Definisi 2.3

Jika $\boldsymbol{u} = (u_1, u_2, \dots, u_n)$ dan $\boldsymbol{v} = (v_1, v_2, \dots, v_n)$ adalah vektor di $\mathbb{R}^n$, maka **hasil kali dalam Eulid** $\boldsymbol{u} \cdot \boldsymbol{v}$ didefinisikan dengan :

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \qquad (2.4)$$

(Anton, 2004, hal: 261)

Selanjutnya, akan dijelaskan tentang $norm$ vektor, $norm$ matriks serta $nullity$ dan $rank$ matriks.

Definisi 2.4

***norm* vektor** di $\mathbb{R}^n$ merupakan fungsi $\|\cdot\|$, dari $\mathbb{R}^n$ ke $\mathbb{R}$ yang memenuhi sifat :

(a) $\|\boldsymbol{x}\| \geq 0$ untuk setiap $\boldsymbol{x} \in \mathbb{R}^n$;

(b) $\|\boldsymbol{x}\| = 0$ jika dan hanya jika $\boldsymbol{x} = \mathbf{0}$;

(c) $\|\alpha \boldsymbol{x}\| = |\alpha| \|\boldsymbol{x}\|$ untuk setiap $\alpha \in \mathbb{R}$ dan $\boldsymbol{x} \in \mathbb{R}^n$;

(d) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ untuk setiap $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$.

(Burden, 2011, hal: 432).

Jarak antara dua vektor dinotasikan dengan $\|\boldsymbol{x}-\boldsymbol{y}\|$ dengan menggunakan definisi norm yang bersesuaian(Burden, 2011, hal: 435). Kelas *norm* vektor yang biasa digunakan adalah *p-norm* .

Definisi 2.5

Misalkan $\boldsymbol{x}$ adalah vektor dalam ruang-n, maka kelas *p-norm* didefinisikan dengan :

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}, p \geq 1 \qquad (2.5)$$

(Golub,1996, hal: 53)

Misalkan $\boldsymbol{x} \in \mathbb{R}^n$, beberapa contoh *p-norm* yang digunakan diantaranya :

$$p = 1 \Rightarrow \|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|\,, \qquad (2.6)$$

$$p = 2 \Rightarrow \|\boldsymbol{x}\|_2 = (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)^{\frac{1}{2}}$$

$$= \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}\,, \qquad (2.7)$$

Persamaan (2.6) disebut juga dengan *Euclidian norm* (Anton, 2005, hal: 263).

Untuk ∞-*norm* didefinisikan dengan :

$$\|\boldsymbol{x}\|_\infty = max_{1 \leq i \leq n} |x_i| \qquad (2.8)$$

Definisi 2.6

Suatu vektor $\boldsymbol{x}$ dinamakan **vektor satuan** yang bersesuaian dengan $\|\cdot\|$ jika memenuhi $\|\boldsymbol{x}\| = 1$. (Golub, 1996, hal: 53)

Definisi 2.7

***Norm* matrix** dalam himpunan matriks-matriks berukuran mxn merupakan fungsi bernilai riil, $\|\cdot\|$, dari $\mathbb{R}^{mxn}$ ke $\mathbb{R}$ dengan $\alpha \in \mathbb{R}$ adalah suatu konstanta, maka berlaku :

(a) $\|A\| \geq 0$ untuk setiap $A \in \mathbb{R}^{mxn}$;

(b) $\|A\| = 0$ jika dan hanya jika $A = O$, yaitu matriks dengan semua anggotanya 0;

(c) $\|\alpha A\| = |\alpha|\|A\|$;

(d) $\|A + B\| \leq \|A\| + \|B\|$;

(Golub, 1996, hal: 55)

Hubungan antara tiga buah *norm* yang berbeda, didefinisikan dengan:

Definisi 2.8

misalkan $\|\cdot\|_\alpha : \mathbb{R}^{mxn} \to \mathbb{R}$, $\|\cdot\|_\beta : \mathbb{R}^{mxq} \to \mathbb{R}$ dan $\|\cdot\|_\gamma : \mathbb{R}^{qxn} \to \mathbb{R}$. Dengan dua matriks A_{mxq} dan B_{qxn}, maka :

$$\|AB\|_\alpha \leq \|A\|_\beta \|B\|_\gamma \qquad (2.9)$$

(Golub, 1996, hal: 55).

Jarak antara dua matriks A dan B berukuran sama didefinisikan dengan $\|A - B\|$ (Burden, 2011, hal: 438). Salah satu definisi *norm* matriks yang umum digunakan adalah *norm Forbenius*.

Teorema 2.9

Norm Forbenius didefinisikan dengan :

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \qquad (2.10)$$

Dan ***p-norm***, didefinisikan dengan :

$$\|A\|_p = sup_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p} \qquad (2.11)$$

Merupakan *norm* matriks (Golub,1996, hal: 55).

Matriks sembarang memiliki sifat yang penting diantaranya *nullity* dan *rank* didefinisikan dengan :

Definisi 2.10

(a) ***Rank*** dari matriks A, dinotasikan dengan $rank(A)$ menunjukkan jumlah baris yang bebas secara linear di A.

(b) ***Nullity*** dari matriks A, dinotasikan dengan $nullity(A)$, merupakan $n - rank(A)$, dan merupakan ukuran terbesar dari himpunan vektor bebas secara linear $\boldsymbol{v} \in \mathbb{R}^n$ dimana $A\boldsymbol{v} = \boldsymbol{0}$.

(Burden, 2011, hal: 614)

2.5 Nilai Eigen dan Vektor Eigen

2.5.1 Definisi Nilai Eigen dan Vektor Eigen

Sebelum mendefinisikan nilai Eigen, diperlukan definisi polinomial karaktersitik yang berguna untuk mencari nilai eigen.

Definisi 2.11

Jika A adalah matriks persegi, **polinomial karakteristik** dari A didefinisikan dengan:

$$\rho(\lambda) = det(A - \lambda I) \qquad (2.12)$$

(Burden, 2011, hal: 443)

Definisi 2.12

Jika ρ merupakan polinomial karakteristik dari matriks persegi A, solusi $\rho(\lambda) = 0$ merupakan **nilai eigen** (***eigen value***)**,** atau nilai karakteristik dari matriks A. Jika λ merupakan nilai eigen tak nol dari A dan $\boldsymbol{x} \neq \mathbf{0}$ memenuhi $(A - \lambda I)\boldsymbol{x} = \mathbf{0}$, maka $\boldsymbol{x}$ merupakan **vektor eigen** (***eigen vector***) atau vektor karakteristik dari A yang bersesuaian dengan nilai eigen λ. (Burden, 2011, hal: 443)

Hubungan antara nilai eigen dengan himpunan vektor yang bebas linear dijelaskan dalam teoreman berikut.

Teorema 2.13

Jika A merupakan matriks dan $\lambda_1, \lambda_2, \dots, \lambda_k$ merupakan k nilai eigen yang berbeda dari A bersesuaian dengn vektor eigen $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x}_k$, maka $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x}_k\}$ adalah himpunan vektor yang bebas linear (Burden, 2011, hal: 565).

2.5.2 Aproksimasi Nilai Eigen

Untuk mendapatkan nilai eigen dan vektor eigen menggunakan definisi 2.12, membutuhkan dua proses terhadap matriks awal yaitu mencari determinan matriks $(A - \lambda I)$ dan mencari solusi dari polinomial karakteristik $\rho(\lambda) = 0$. Secara komputasi, Mencari nilai eigen dan vektor eigen dengan kedua proses ini membutuhkan sumber daya besar. Mencari determinan matriks membutuhkan biaya komputasi yang besar. Selain itu, untuk mendapatkan aproksimasi terbaik solusi $\rho(\lambda) = 0$ juga sulit (Burden, 2011, hal: 443). Oleh karena itu, dibutuhkan suatu algoritma yang dapat menentukan nilai eigen dari suatu matriks secara lebih efektif. Pada umumnya, metode yang digunakan untuk mengaproksimasi nilai eigen menggunakan faktorisasi matriks.

2.6 Ortogonalitas dan Proses *Gram-Schmidt*

Singular value decomposition (SVD) dari suatu matriks A membentuk matriks ortogonal dan matriks diagonal, maka diperlukan definisi vektor ortonormal yang menyusun matriks ortogonal. Metode yang digunakan untuk membentuk himpunan vektor ortogonal adalah proses gram-schmidt.

Definisi 2.14

Himpunan vektor kolom $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_n}\}$ disebut **ortogonal** jika $\boldsymbol{v_i} \cdot \boldsymbol{v_j} = \boldsymbol{v_i^T v_j} = 0$, untuk setiap $i \neq j$. Selanjutnya, jika $\boldsymbol{v_i} \cdot \boldsymbol{v_i} = \boldsymbol{v_i^T v_i} = 1$, untuk setiap $i = 1,2, \dots, n$, maka himpunan tersebut dikatakan **ortonormal**. (Burden, 2011, hal: 566)

Proses **Gram-Schmidt** dapat digunakan untuk membentuk himpunan vektor ortogonal dari himpunan vektor yang bebas secara linear.

Teorema 2.15 (Proses Gram-Schmidt)

Misalkan $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}\}$ merupakan himpunan k vektor di $\mathbb{R}^n$. Maka $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots, \boldsymbol{v_k}\}$ yang didefinisikan dengan :

$\boldsymbol{v_1} = \boldsymbol{x_1}$;

$\boldsymbol{v_2} = \boldsymbol{x_2} - \left(\frac{\boldsymbol{v_1^T x_2}}{\boldsymbol{v_1^T v_1}}\right)\boldsymbol{v_1}$;

$\boldsymbol{v_3} = \boldsymbol{x_3} - \left(\frac{\boldsymbol{v_1^T x_3}}{\boldsymbol{v_1^T v_1}}\right)\boldsymbol{v_1} - \left(\frac{\boldsymbol{v_2^T x_3}}{\boldsymbol{v_2^T v_2}}\right)\boldsymbol{v_2}$;

$\vdots$

$\boldsymbol{v_k} = \boldsymbol{x_k} - \sum_{i=1}^{k-1}\left(\frac{\boldsymbol{v_i^T x_k}}{\boldsymbol{v_i^T v_i}}\right)\boldsymbol{v_i}$.

merupakan himpunan k vektor yang ortogonal di $\mathbb{R}^n$. (Burden, 2011, hal: 567)

Definisi 2.16

Matriks persegi A disebut **matriks ortogonal** jika setiap vektor kolom dari matriks A, yaitu $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \dots \boldsymbol{v_n}\}$ merupakan himpunan matriks ortonormal di $\mathbb{R}^n$ (Burden, 2011, hal: 570).

Beberapa sifat dari matriks ortogonal disebutkan dalam teorema berikut.

Teorema 2.17

Misalkan Q merupakan matriks ortogonal berukuran nxn, maka :

(a) Q memiliki *inverse* dengan $Q^{-1} = Q^T$;

(b) Untuk sembarang $\boldsymbol{x}$ dan $\boldsymbol{y}$ di $\mathbb{R}^n$, berlaku $(Q\boldsymbol{x})^T Q\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y}$;

(c) Untuk sembarang $\boldsymbol{x}$ di $\mathbb{R}^n$, berlaku $\|Q\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$.

(Burden, 2011, hal: 570)

2.7 Diagonalisasi Ortogonal

Salah satu tujuan faktorisasi matriks yang disebutkan pada subbab 2.3 adalah membentuk matriks yang lebih sederhana, dan atau memiliki jumlah komponen yang lebih sedikit. Salah satu hasil dari faktorisasi matriks adalah matriks diagonal dimana setiap komponennya bernilai 0 kecuali komponen pada diagonal utamanya. Pembentukan matriks diagonal diberikan pada definisi di bawah ini.

Definisi 2.18

Suatu matriks bujur sangkar A dikatakan **dapat didiagonalkan** jika ada suatu matriks yang dapat dibalik P sedemikian sehingga $P^{-1}AP$ adalah suatu matriks diagonal. Matriks P dikatakan **mendiagonalkan** A (Anton, 2004, hal: 552).

Teorema 2.19

Jika A suatu matriks persegi, maka pernyataan berikut ekuivalen:

(a) A dapat didiagonalkan;

(b) A memiliki n vektor eigen yang bebas secara linear.

(Anton, 2004, hal: 552).

Selanjutnya dengan sifat khusus dari matriks simetri, didapat teorema yang menunjukkan bahwa matriks P mendiagonalkan A secara ortogonal.

Teorema 2.20

Matriks A adalah simetri jika dan hanya jika terdapat matriks diagonal D dan matriks ortogonal P dengan $A = PDP^T$ (Burden, 2011, hal : 572).

2.8 *Singular Value Decomposition* (SVD)

2.8.1 Definisi SVD

Singular value decomposition (SVD) merupakan salah satu faktorisasi matriks dengan *approximation method*. SVD memanfaatkan sifat matriks simetri yang dibentuk dari operasi terhadap matriks sembarang A berukuran mxn. Sifat tersebut dijamin oleh teorema di bawah ini.

Teorema 2.21

Misalkan A matriks berukuran mxn, maka berlaku:

(a) Matriks A^TA dan AA^T simetri;

(b) $nullity(A) = nullity(A^TA)$;

(c) $rank(A) = rank(A^TA)$;

(d) nilai eigen dari A^TA dan AA^T adalah riil dan tak negatif ;

(e) nilai eigen tak nol dari AA^T dan A^TA sama.

(Burden, 2011, hal: 614)

Selanjutnya, dengan menggunakan teorem 2.20 dan definisi nilai *singular* di bawah ini, maka dapat dibentuk suatu faktorisasi SVD pada teorema 2.23.

Definisi 2.22

Nilai *singular* dari matriks A berukuran mxn adalah akar kuadrat positif dari nilai eigen tak nol matriks simetri A^TA. (Burden, 2011).

Didapat bentuk persamaan SVD sebagai berikut.

Teorema 2.23

(*Singular Value Decomposition* (SVD)) jika A merupakan matriks real berukuran mxn maka terdapat matriks-matriks ortogonal :

$S = [\boldsymbol{s}_1, \dots, \boldsymbol{s}_m] \in \mathbb{R}^{mxm}$ dan $D = [\boldsymbol{d}_1, \dots, \boldsymbol{d}_n] \in \mathbb{R}^{nxn}$

sedemikian sehigga :

$$S^TAD = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{mxn}, \quad p = \min\{m, n\} \qquad (2.13)$$

Dimana σ_i adalah nilai *singular* dari A dengan $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. (Golub:1996, hal: 70)

Bentuk lain dari persamaan (2.13) adalah :

$$A = S\Sigma D^T \qquad (2.14)$$

2.8.2 Algoritma SVD

Dengan tanpa menghilangkan keumuman, algoritma berikut ini menunjukkan *singular value decomposition* dari matriks A berukuran mxn dengan $m \geq n$.

Algoritma SVD

Input **:** matriks A berukuran mxn

Output **:** matriks ortogonal S berukuran mxm,
matriks semidiagonal Σ berukuran mxn berisi nilai *singular* dari A
matriks ortogonal D^T berukuran nxn
sedemikian sehingga berlaku $A = S\Sigma D^T$

STEP 1 hitung matriks A^T.

STEP 2 hitung matriks simetri $A^T A$.

STEP 3 hitung semua nilai eigen dari matriks simetri $A^T A$.

STEP 4 urutkan nilai-nilai eigen dari $A^T A$ sedemikian sehingga berlaku :

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_k^2 > \sigma_{k+1} = \cdots = \sigma_n = 0$$

dimana σ_k^2 merupakan nilai eigen tak nol terkecil dari $A^T A$

STEP 5 dengan definisi 2.22, $\sigma_i, i = 1, \ldots, n$ merupakan nilai-nilai *singular* yang menjadi elemen diagonal dari Σ.

STEP 6 faktorisasi $A^T A = DMD^T$, dengan $M = \Sigma^2$ merupakan matriks diagonal dengan setiap elemennya merupakan nilai eigen dari $A^T A$. D merupakan matriks ortogonal dengan setiap vektor kolom dari D, merupakan vektor eigen yang bersesuain dengan nilai eigen pada elemen diagonal M. Didapat matriks $D = [\boldsymbol{d}_1 \boldsymbol{d}_2 \ldots \boldsymbol{d}_m]$.

STEP 7 vektor kolom dari S yang bersesuaian dengan nilai *singular* tak nol σ_i didefinisikan dengan :

$$\boldsymbol{s}_i = \frac{1}{\sigma_i} A\boldsymbol{d}_i \quad , i = 1,2, \ldots, k$$

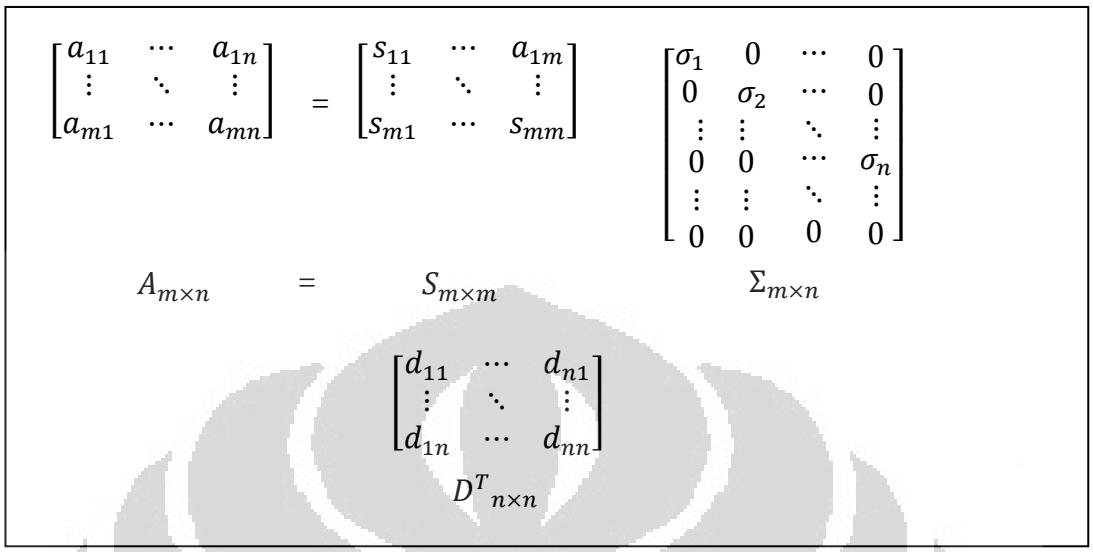
sebagai k-vektor pertama dari S.

STEP 8 gunakan proses *gram-schmidt* untuk vektor kolom $\boldsymbol{s}_{k+1}, \boldsymbol{s}_{k+2}, \ldots, \boldsymbol{s}_m$ pada S.

STEP 9 Susun $S = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m]$

STEP 10 OUTPUT (S, Σ, D^T)

Keluaran algoritma SVD di atas dapat diilustrasikan dengan gambar di bawah ini.



Gambar 2.1 Ilustrasi SVD

2.8.3 Pengecilan SVD

Menurut Burden (2011) SVD dapat digunakan dalam banyak aplikasi karena SVD mampu memberikan fitur-fitur penting dalam matriks A berukuran mxn dengan menggunakan ukuran yang jauh lebih kecil. Nilai *singular* yang disusun menurun pada matriks semidiagonal Σ. Dengan menyisakan k baris dari matriks Σ memberikan pendekatan yang baik terhadap matriks A.

Teorema berikut menunjukkan jarak antara matriks awal A dengan matriks pendekatan $\tilde{A}_k$ dapat dihitung dengan baik menggunakan nilai-nilai *singular*.

Teorema 2.24

Misalkan SVD A diberikan pada teorema 2.23, dengan $k < r = rank(A)$ dan

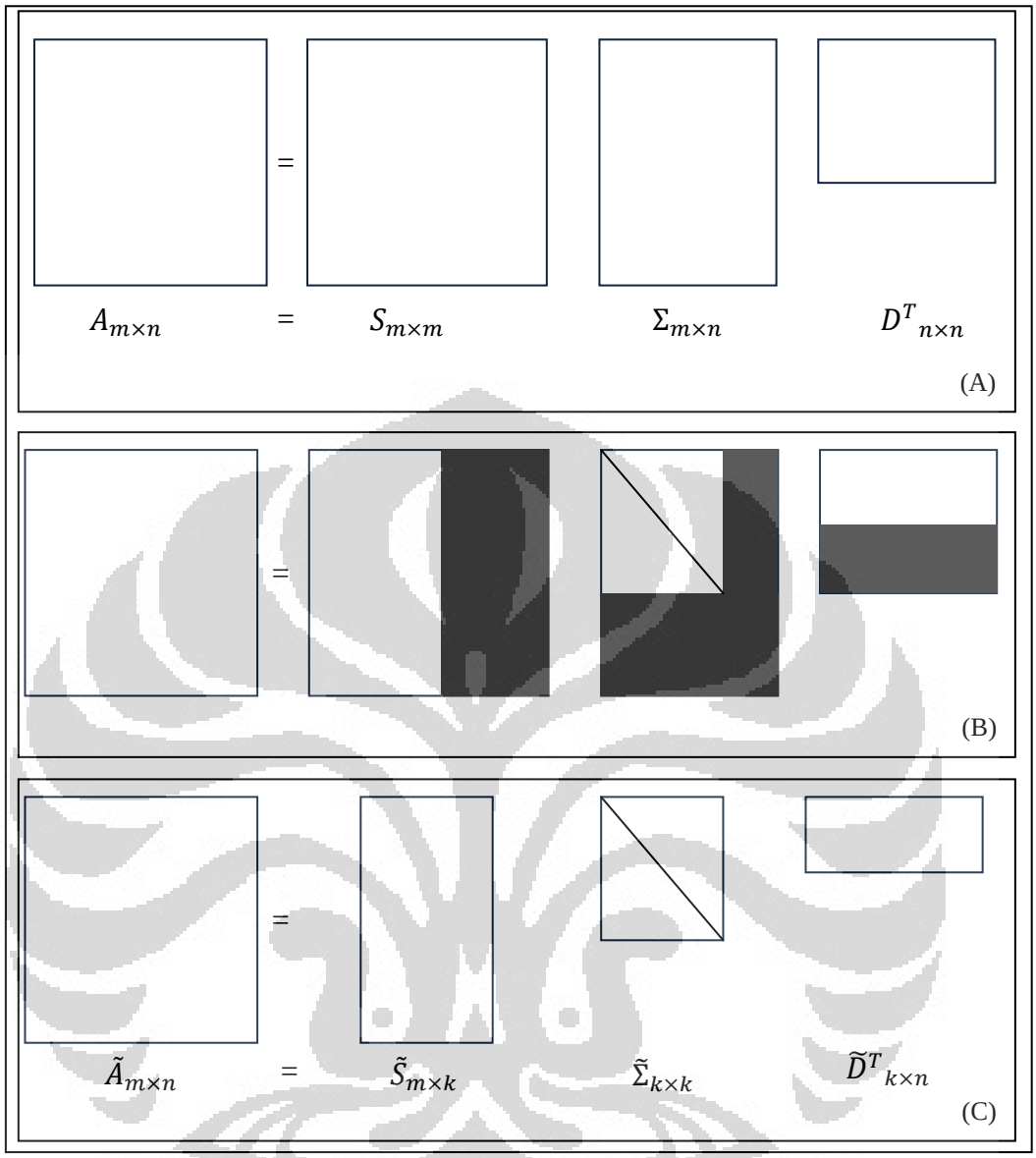
$$\tilde{A}_k = \tilde{S}_{mxk}\tilde{\Sigma}_{kxk}\tilde{D}^T_{kxn} \quad (2.15)$$

maka

$$\left\|A - \tilde{A}_k\right\|_F = \sqrt{\sigma^2_{k+1} + \cdots + \sigma^2_p} \quad (2.16)$$

(Golub:1996, hal : 72)

Pengecilan SVD diilustrasi dengan gambar berikut ini.



Keterangan : (A) = SVD dari matriks *A* berukuran mxn
(B) = Proses pengecilan SVD
(C) = Hasil pengecilan SVD

Gambar 2.2 Ilustrasi pengecilan SVD

Gambar 2.2 memberikan gambaran bagaimana proses pengecilan SVD. Dengan memiliki suatu matriks A berukuran mxn, didapat suatu faktorisasi SVD yang memenuhi persamaan 2.14 pada gambar (A). Selanjutnya, dengan memilih nilai k yang memenuhi $k < r = rank(A)$, lakukan hal-hal berikut :

(a) Bentuk matriks $\tilde{S}_{mxk}$ dengan menghapus vektor-vektor kolom $\{\boldsymbol{s}_{k+1}, \dots, \boldsymbol{s}_m\}$ pada matriks S;

(b) Bentuk matriks $\tilde{\Sigma}_{kxk}$ dengan mengambil k nilai *singular* dari matriks Σ;

(c) Bentuk matriks $\widetilde{D^T}_{kxn}$ dengan menghapus vektor-vektor baris $\{\boldsymbol{d}^T{}_{k+1}, \ldots, \boldsymbol{d}^T{}_n\}$ pada matriks D^T.

Langkah (a), (b) dan (c) diilustrasikan pada gambar (B). Gambar (C) menunjukkan hasil pengurangan SVD terhadap matriks A menjadi $\tilde{A}$.

BAB 3
PENGGUNAAN *LATENT SEMANTIC ANALYSIS* (LSA) UNTUK EKSTRAKSI TOPIK UTAMA

Pada bab ini akan diberikan penjelasan bagaimana mengekstraksi topik utama dari kumpulan dokumen berita dengan menggunakan metode *latent semantic analysis* (LSA). Dimulai dengan pengertian LSA, kemudian dilanjutkan dengan proses akuisis data, pembentukan matriks kata-dokumen dan ekstraksi topik utama. Pada bagian akhir bab ini diberikan satu contoh sederhana metode LSA untuk mengekstraksi topik utama.

3.1 *Latent Semantic Analysis* (LSA)

Metode LSA merupakan metode matematis untuk mengekstraksi dan mendapatkan hubungan konteks penggunaan kata dalam paragraf acak (Launder, 1998). LSA merupakan bagian dari *machine learning* yang termasuk dalam *unsupervised learning.*

Langkah awal metode LSA ini adalah merepresentasikan hubungan kumpulan kata dengan kumpulan dokumen kedalam suatu matriks. Baris pada matriks merepresentasikan kata yang berbeda dan kolom pada matriks merepresentasikan dokumen. Setiap anggota matriks menunjukkan frekuensi penggunaan kata dalam dokumen. Matriks hubungan kata dengan dokumen ini disebut sebagai matriks kata-dokumen.

Selanjutnya, faktorisasi SVD diimplemantasikan pada matriks kata-dokumen untuk ekstraksi hubungan semantik tersembunyi. Dengan SVD, kata dan dokumen yang memiliki kedekatan semantik di letakkan berdekatan. Pengelompokkan data dengan SVD diurutkan dari kelompok yang memiliki ‘pengaruh’ terbesar hingga terkecil. Karena vektor-vektor disusun berdasarkan urutan nilai *singular*.

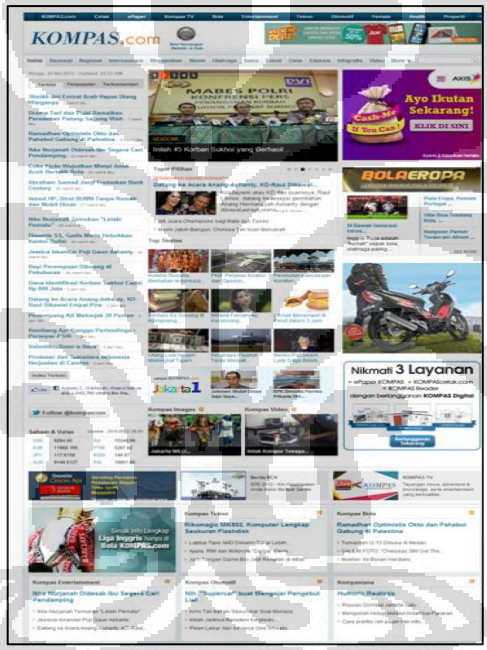
3.2 Akuisisi Data

Langkah awal untuk menggunakan metode LSA adalah penentuan sumber data. Mengekstraksi topik utama harian berita *online* nasional membutuhkan

dokumen berita *online* yang didapatkan melalui protal berita *online.* Portal tersebut juga menyediakan RSS *feed* sebagai ringkasan dan *update* dokumen berita terbaru.

3.2.1 Portal Berita *Online*

Portal berita online merupakan suatu alamat *web* yang dimiliki oleh kantor berita. Melalui portal inilah kantor berita menyajikan berita dalam bentuk dokumen *digital* yang dapat diakses melalui internet. Gambar berikut merupakan salah satu contoh halaman awal portal berita *online* nasional.



[sumber : http://www.kompas.com pada tanggal 20 Mei 2012 pk. 22.20]

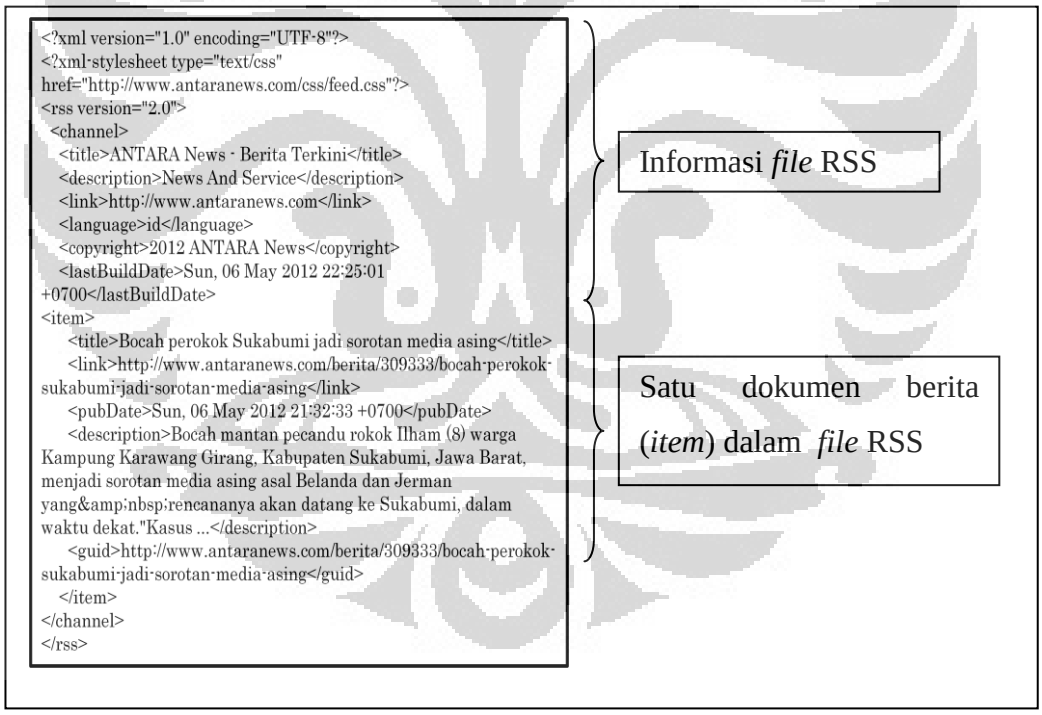
Gambar 3.1 Tampilan halaman awal (*homepage*) www.kompas.com

Gambar di atas, memperlihatkan halaman awal portal berita Kompas. Halaman awal ini menyajikan berbagai hal dimulai dari berita utama (*head line*), berita pilihan hingga iklan. Tampilan awal ini pada umumnya dibangun semenarik mungkin untuk memberikan kenyamanan dan informasi bagi pembaca.

3.2.2 RSS

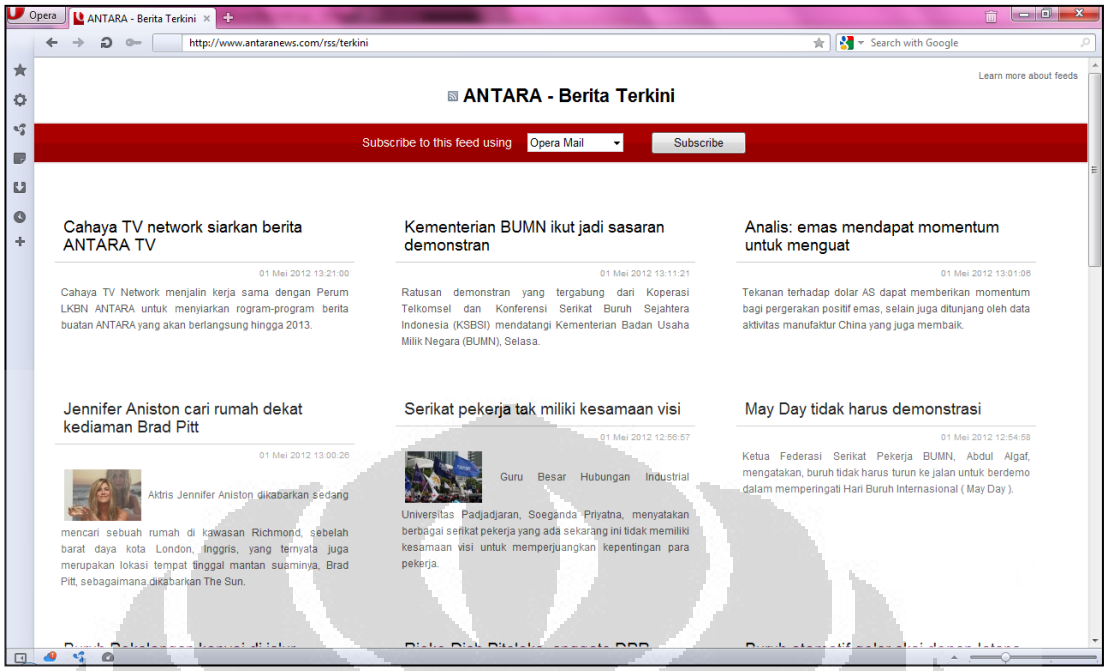
RSS merupakan singkatan dari *Really Simple Syndication* (RSS 2.0) atau *Rich Site Summary* (RSS 0.91). RSS adalah suatu *file* berformat XML (*extensible markup language*). XML merupakan *file* yang digunakan untuk membuat *markup* untuk keperluan pertukaran data antar sistem yang beraneka ragam dan merupakan kelanjutan dari HTML (*hyper text markup language*). Selanjutnya, RSS digunakan untuk sindikasi portal situs berita dan *weblog*. Dengan menggunakan RSS, pengguna internet dapat berlangganan kepada situs web yang menyediakan umpan *web* (*feed*) RSS.

Diantara informasi yang berguna dalam *file* RSS adalah versi RSS, *encoding*, alamat situs sumber RSS, hak cipta, tanggal pembuatan dan berita terbaru yang dipublikasikan di dalam situs utama. *File* RSS mengelompokkan setiap dokumen dan informasinya dalam sebuah kelompok *item.*



Gambar 3.2 Contoh isi file RSS dan keterangan

Setiap *item* berita memiliki informasi berupa judul berita, alamat situs berita, waktu penerbitan berita, serta deskripsi berupa potongan berita. Gambar berikut ini menampilkan hasil kompilasi *file* RSS dari portal berita ANTARA.



Gambar 3.3 Tampilan kompilasi RSS dengan *web browser* Opera**.**

Berbeda dengan *file* HTML biasa pada *hompage* portal berita pada (gambar 3.1) , *file* RSS (gambar 3.2 dan gambar 3.3) hanya berisikan informasi berupa berita terkini dengan informasi berupa judul dan potongan berita.

3.3 Penyiapan Data

Selanjutnya, dokumen berita yang telah dikumpulkan Tahapan dalam pengumpulan kata ini dijelaskan oleh Murfi (2010), terdiri dari langkah-langkah terurut berikut ini:

(a) Penghapusan *Tag* HTML atau Format

Sumber data yang digunakan berupa *file* RSS (gambar 3.2) masih memiliki *tag* HTML. Untuk mendapatkan dokumen yang akan diolah terdiri dari judul dan kalimat berita (gambar 3.3) maka *tag* HTML perlu dihapuskan.

(b) *Tokenization*

Tokenization merupakan tahap pengolahan dokumen untuk menghilangkan gambar, tanda baca dan karakter selain huruf (*non-alpha character*). Tahap ini memecah dokumen menjadi struktur terkecil yang dapat diproses, yaitu kata. Setelah melewati proses ini, dokumen berubah menjadi barisan kata dengan huruf kecil.

(c) *Filtering*

Filtering merupakan proses pemilihan kata-kata yang akan digunakan untuk merepresentasikan dokumen sehingga dapat digunakan untuk :

- Mendeskripsikan isi dokumen.
- Membedakan dokumen dengan dokumen lain dalam kumpulan dokumen.

Kata-kata yang mungkin tidak memiliki makna yang berarti ketika berdiri sendiri, seperti kata hubung, kata depan dan negasi dihilangkan dari dari setiap dokumen.

(d) *Weighting*

Weighting atau pembobotan berguna untuk menunjukkan hubungan kata dengan dokumen. Setiap hubungan kata dengan dokumen direpresentasikan dengan bilangan.

Dokumen $\boldsymbol{d}$ direpresentasikan sebagai vektor $\boldsymbol{d} = (t_1, t_2, \dots, t_n)$, dimana t_j merupakan kata ke-j di $\boldsymbol{d}$ dan n adalah jumlah kata yang berbeda dalam $\boldsymbol{d}$. Selanjutnya, pembobotan merupakan suatu fungsi yang didefinisikan dengan:

$$f_j(\boldsymbol{d}) = k_j \qquad (3.1)$$

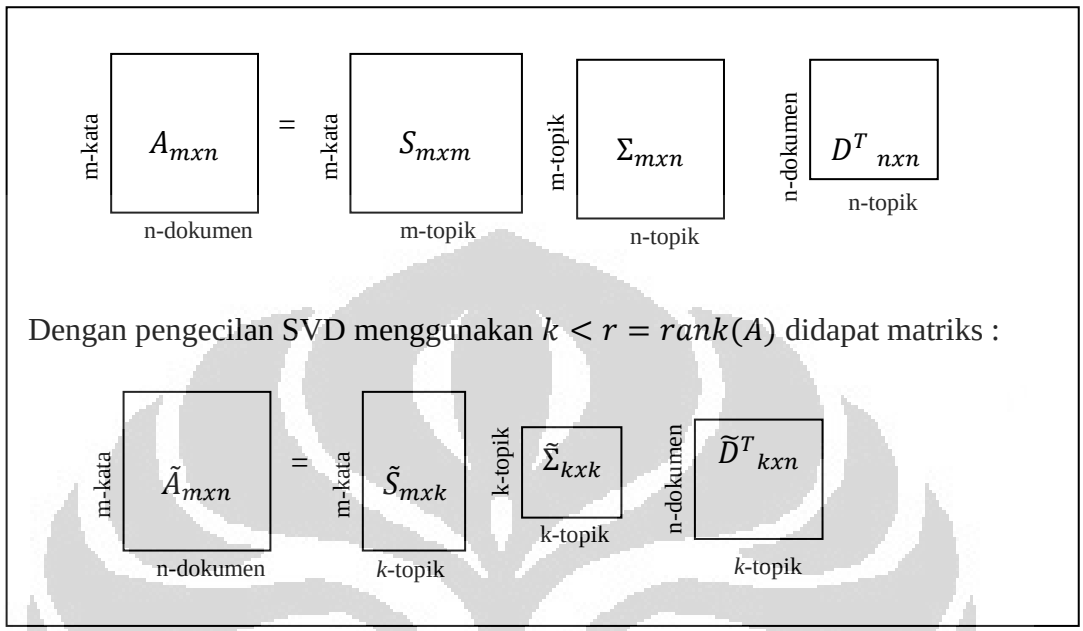
Dimana f_j merupakan fungsi untuk kata t_j di $\boldsymbol{d}$ dan k_j adalah banyaknya kata t_j di dalam dokumen $\boldsymbol{d}$.

Kumpulan data yang sudah melalui tahap penyiapan (a), (b), (c) dan (d) kemudian dibentuk menjadi suatu matriks kata-dokumen. Baris matriks merepresentasikan kata-kata yang berbeda dari seluruh dokumen berita. Kolom matriks merupakan representasi dari dokumen berita. Setiap anggota matriks menunjukkan jumlah kata dalam dokumen. Oleh karena itu, ukuran matriks menunjukkan banyaknya kata dan dokumen yang digunakan dalam proses LSA.

3.4 Ekstraksi Topik Utama dengan LSA

Tujuan utama dari proses LSA ini adalah untuk mengekstrak suatu hubungan semantik tersembunyi dalam kumpulan dokumen berita. Hubungan semantik yang dimaksud dalam tugas akhir ini adalah topik utama.

Dengan menggunakan SVD, dan dilanjutkan dengan pengurangan SVD menjadi $rank(k)$ dengan k menunjukkan jumlah topik yang akan diinterpretasikan, maka akan didapat ilustrasi pada gambar di bawah ini.



Gambar 3.4 SVD matriks kata-dokumen pada LSA

Pada gambar 3.4 terlihat bahwa matriks kata-dokumen di faktorisasi SVD menjadi tiga buah matriks. Dengan pengurangan SVD, didapat matriks aproksimasi yang menginterpretasikan hubungan baru antara kata dengan dokumen. Interpretasi pengurangan SVD pada LSA dijelaskan pada tabel di bawah ini.

Tabel 3.1 Interpretasi komponen SVD dalam LSA

$\tilde{A}_{m\times n}$	: aproksimasi terbaik $rank(k)$ untuk matriks A
$\tilde{S}_{m\times k}$	: Matriks hubungan antara kata dengan topik
$\tilde{\Sigma}_{k\times k}$	: Nilai *singular*,
$\tilde{D}^T{}_{k\times n}$	: Matriks hubungan antara dokumen dengan topik
m	: banyak kata
n	: banyak dokumen
k	: banyak topik

[sumber : Bery, Dumais, O'brien, 1994, hal:4]

Dengan SVD dan pengecilan SVD, didapat matriks aproksimasi $\tilde{S}_{mxk}$ yang merepresentasikan hubungan kata dengan topik. Sehingga dengan mengurutkan nilai dari terbesar hingga terkecil dari setiap k-vektor kolom, didapatkan urutan kata yang bersesuaian dengan indeks elemen vektor yang terurut.

3.5 Contoh Ekstraksi Topik Utama dengan LSA

Dengan menggunakan metode LSA, akan ditunjukkan pencarian topik utama dari tujuh dokumen dan tujuh belas kata yang dilibatkan. Tabel berikut berisi judul dokumen berita yang digunakan sebagai sumber data.

Tabel 3.2 Contoh kumpulan dokumen

No	Dokumen	Sumber
1	BK DPR Kaji Aturan Senpi di Kode Etik Anggota Dewan.	Detik.com, 08 Mei 2012 pk. 11:00:33 WIB
2	Keributan Pecah Dinihari, 7-Eleven Salemba Diberi Garis Polisi.	Detik.com, 08 Mei 2012 pk. 11:44:26 WIB
3	Keributan di 7-Eleven Salemba Menyisakan Bercak Darah.	Detik.com, 08 Mei 2012 pk. 11:56:35 WIB
4	Pelaku Keributan di 7-Eleven Salemba Diduga 20 Orang.	Detik.com, 08 Mei 2012 pk. 12:09:30 WIB
5	Keributan di 7-Eleven Salemba, 1 Orang Terluka.	Detik.com, 08 Mei 2012 pk. 12:20:32 WIB
6	Anggota DPR Harus Kembalikan Senjata Apinya ke Polisi.	Detik.com, 07 Mei 2012 10:10:11 WIB
7	Aturan Kepemilikan Senpi Sudah Baik, Tapi Penegakannya Lemah.	Detik.com, 07 Mei 2012 11:25:10 WIB

Setelah didapatkan sumber data, tahapan selanjutnya adalah penyiapan data melalui *tokenization, filtering* dan *weighting.* Setiap proses digambarkan dalam tabel 3.3. Pada contoh ini, proses *filtering* dilakukan secara manual untuk mendapatkan kata-kata yang digunakan sesuai dengan tujuan pada subbab 3.3 bagian (c).

Tabel 3.3 Proses penyiapan data

No.	*Tokenization*	*Filtering*	*Weighting*
1	bk, dpr, kaji, aturan, senpi, di, kode, etik, anggota, dewan	[dpr, aturan, senpi, anggota]	{ dpr : 1, aturan : 1, senpi : 1, anggota : 1}
2	keributan, pecah, dinihari, eleven, salemba, diberi, garis, polisi	[keributan, eleven, salemba, polisi]	{ keributan : 1, eleven : 1, salemba : 1, polisi : 1}
3	keributan, di, eleven, salemba, menyisakan, bercak, darah	[keributan, eleven, salemba]	{ keributan : 1, eleven : 1, salemba : 1}
4	pelaku, keributan, di, eleven, salemba, diduga, orang	[pelaku, keributan, eleven, salemba]	{ pelaku : 1 , keributan : 1, eleven : 1, salemba : 1 }
5	keributan, di, eleven, salemba, orang, terluka	[keributan, eleven, salemba, terluka]	{Keributan : 1, eleven : 1, salemba : 1, terluka : 1}
6	anggota, dpr, harus, kembalikan, senjata, apinya, ke, polisi	[anggota, dpr, senjata, polisi]	{anggota : 1 , dpr : 1 , senjata : 1 , polisi : 1}
7	aturan, kepemilikan, senpi, sudah, baik, tapi, penegakannya, lemah	[aturan, kepemilikan, senpi]	{ aturan : 1, kepemilikan : 1 , senpi : 1}

Dengan menggunakan kumpulan kata yang digunakan dalam proses LSA, dibangun sebuah *dictionary* atau kamus berisikan kata-kata yang berbeda. Dengan bantuan kamus tersebut, dapat dibangun hubungan kata dan dokumen yang dijelaskan pada tabel berikut.

Tabel 3.4 Hubungan kata dan dokumen

		Dokumen						
		d1	d2	d3	d4	d5	d6	d7
Kata	dpr	1	0	0	0	0	1	0
	aturan	1	0	0	0	0	0	1
	senpi	1	0	0	0	0	0	1
	anggota	1	0	0	0	0	1	0
	keributan	0	1	1	1	1	0	0
	eleven	0	1	1	1	1	0	0
	salemba	0	1	1	1	1	0	0
	polisi	0	1	0	0	0	1	0
	pelaku	0	0	0	1	0	0	0
	terluka	0	0	0	0	1	0	0
	senjata	0	0	0	0	0	1	0
	kepemilikan	0	0	0	0	0	0	1

Dari tabel di atas kemudian dibentuk sebuah matriks kata-dokumen A berukuran $12x7$ dengan setiap anggotanya menunjukkan frekuensi kata dalam dokumen.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Selanjutnya, algoritma SVD diimplementasikan terhadap matriks A, dengan persamaan $A = S\Sigma D^T$ dimana matriks S berukuran $12x12$, matriks Σ berukuran $12x7$ serta matriks D^T berukuran $7x7$.

$$S = \begin{bmatrix} -0.0213 & -0.4972 & -0.2528 & -0.3112 & 0.0000 & 0.0498 & 0.0764 & -0.0641 & 0.6625 & 0.1381 & -0.2079 & 0.2794 \\ -0.0049 & -0.4324 & 0.4525 & 0.0454 & -0.0000 & 0.1035 & -0.1104 & -0.6755 & 0.0103 & -0.1218 & 0.2031 & -0.2661 \\ -0.0049 & -0.4324 & 0.4525 & 0.0454 & -0.0000 & 0.1035 & -0.1104 & 0.7268 & 0.0598 & 0.0114 & 0.1633 & -0.1574 \\ -0.0213 & -0.4972 & -0.2528 & -0.3112 & 0.0000 & 0.0498 & 0.0764 & 0.0128 & -0.7325 & -0.0276 & -0.1584 & 0.1441 \\ -0.5576 & 0.0369 & 0.0425 & -0.0149 & 0.0000 & 0.0732 & 0.1169 & -0.0572 & -0.0781 & 0.7899 & 0.0628 & -0.1716 \\ -0.5576 & 0.0369 & 0.0425 & -0.0149 & 0.0000 & 0.0732 & 0.1169 & 0.0516 & 0.0705 & -0.4445 & -0.5743 & -0.3629 \\ -0.5576 & 0.0369 & 0.0425 & -0.0149 & 0.0000 & 0.0732 & 0.1169 & 0.0056 & 0.0077 & -0.3454 & 0.5115 & 0.5345 \\ -0.1605 & -0.2367 & -0.4515 & 0.6476 & -0.0000 & 0.1315 & -0.5269 & -0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0000 \\ -0.1418 & 0.0196 & 0.0483 & -0.2434 & -0.7071 & -0.4852 & -0.4272 & -0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0000 \\ -0.1418 & 0.0196 & 0.0483 & -0.2434 & 0.7071 & -0.4852 & -0.4272 & -0.0000 & 0.0000 & 0.0000 & -0.0000 & 0.0000 \\ -0.0172 & -0.2177 & -0.3625 & 0.1381 & 0.0000 & -0.5105 & 0.4559 & 0.0513 & 0.0700 & -0.1105 & 0.3663 & -0.4235 \\ -0.0008 & -0.1530 & 0.3427 & 0.4946 & 0.0000 & -0.4569 & 0.2691 & -0.0513 & -0.0700 & 0.1105 & -0.3663 & 0.4235 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 3.5772 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 2.5794 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.9079 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0878 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0000 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7397 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.4236 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$D^T = \begin{bmatrix} -0.0147 & -0.5125 & -0.4676 & -0.5073 & -0.5073 & -0.0616 & -0.0030 \\ -0.7208 & -0.0488 & 0.0430 & 0.0506 & 0.0506 & -0.5617 & -0.3946 \\ 0.2093 & -0.1699 & 0.0668 & 0.0921 & 0.0921 & -0.6916 & 0.6539 \\ -0.4887 & 0.5543 & -0.0410 & -0.2647 & -0.2647 & 0.1502 & 0.5380 \\ -0.0000 & 0.0000 & 0.0000 & -0.7071 & 0.7071 & -0.0000 & 0.0000 \\ 0.4145 & 0.4749 & 0.2971 & -0.3589 & -0.3589 & -0.3776 & -0.3379 \\ -0.1608 & -0.4163 & 0.8277 & -0.1810 & -0.1810 & 0.1931 & 0.1140 \end{bmatrix}$$

Dengan pengecilan SVD, pada contoh ini dipilih nilai $k = 3$ untuk mendapatkan tiga buah topik. Selanjutnya didapat matriks-matriks aproksimasi berikut :

$$\tilde{S} = \begin{bmatrix} -0.0213 & -0.4972 & -0.2528 \\ -0.0049 & -0.4324 & 0.4525 \\ -0.0049 & -0.4324 & 0.4525 \\ -0.0213 & -0.4972 & -0.2528 \\ -0.5576 & 0.0369 & 0.0425 \\ -0.5576 & 0.0369 & 0.0425 \\ -0.5576 & 0.0369 & 0.0425 \\ -0.1605 & -0.2367 & -0.4515 \\ -0.1418 & 0.0196 & 0.0483 \\ -0.1418 & 0.0196 & 0.0483 \\ -0.0172 & -0.2177 & -0.3625 \\ -0.0008 & -0.1530 & 0.3427 \end{bmatrix}$$

$$\tilde{\Sigma} = \begin{bmatrix} 3.5772 & 0.0 & 0.0 \\ 0.0 & 2.5794 & 0.0 \\ 0.0 & 0.0 & 1.9079 \end{bmatrix}$$

$$\tilde{D}^T = \begin{bmatrix} -0.0147 & -0.5125 & -0.4676 & -0.5073 & -0.5073 & -0.0616 & -0.0030 \\ -0.7208 & -0.0488 & 0.0430 & 0.0506 & 0.0506 & -0.5617 & -0.3946 \\ 0.2093 & -0.1699 & 0.0668 & 0.0921 & 0.0921 & -0.6916 & 0.6539 \end{bmatrix}$$

Dari matriks $\tilde{S}$, dengan tabel 3.1 yang menunjukkan interpretasi SVD pada LSA, didapat hubungan antara kata dengan topik. Maka didapat topik utama dari kumpulan dokumen yang diberikan untuk masing-masing topik.

Tabel 3.5 Hubungan kata dengan topik

		Topik		
		t1	t2	t3
Kata	dpr	-0.0213	-0.4972	-0.2528
	aturan	-0.0049	-0.4324	0.4525
	senpi	-0.0049	-0.4324	0.4525
	anggota	-0.0213	-0.4972	-0.2528
	keributan	-0.5576	0.0369	0.0425
	eleven	-0.5576	0.0369	0.0425
	salemba	-0.5576	0.0369	0.0425
	polisi	-0.1605	-0.2367	-0.4515
	pelaku	-0.1418	0.0196	0.0483
	terluka	-0.1418	0.0196	0.0483
	senjata	-0.0172	-0.2177	-0.3625
	kepemilikan	-0.0008	-0.1530	0.3427

Dengan memilih lima kata pertama yang memiliki nilai terbesar dari masing-masing topik memberikan hasil yang ditunjukkan pada tabel 3.6. Ketiga topik yang terekstraksi adalah:

(a) Aturan kepemilikan senjata api di lingkungan DPR;

(b) Pelaku keributan di 7-eleven terluka;

(c) Aturan kepemilikan senpi dan pelaku yang terluka.

Topik (a) dan (b) merupakan dua topik yang berbeda. Sedangkan topik (c) merupakan gabungan dari topik (a) dan (b). Hal ini terjadi karena dokumen yang digunakan pada tabel 3.2 secara intuisi penulis hanya menunjukkan dua topik yang

berbeda yaitu aturan kepemilikan senjata api di lingkungan DPR (topik (a)) dan keributan yang terjadi di 7-Eleven Salemba (topik (b)). Sehingga dengan mengambil lebih dari dua topik menggunakan LSA, maka topik ketiga dan seterusnya mungkin memberikan hasil berupa :

(a) pengulangan topik sebelumnya;

(b) gabungan dari topi-topik sebelumnya;

(c) tidak dapat diinterpretasikan.

Tabel 3.6 Hasil contoh pencarian topik utama dengan LSA

No	Topik pertama		Topik kedua		Topik ketiga	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	kepemilikan	-0.0008	keributan	0.0369	aturan	0.4525
2	Aturan	-0.0049	eleven	0.0369	senpi	0.4525
3	senpi	-0.0049	salemba	0.0369	kepemilikan	0.3427
4	senjata	-0.0172	pelaku	0.0196	pelaku	0.0483
5	dpr	-0.0213	terluka	0.0196	terluka	0.0483
Topik	Aturan kepemilikan senjata api di lingkungan DPR.		Pelaku keributan di 7-Eleven terluka.		Aturan kepemilikan senpi dan pelaku yang terluka.	

Sebagai proses lanjutan, dengan menghitung selisih dari matriks A dengan matriks aproksimasi $\tilde{A} = \tilde{S}\tilde{\Sigma}\tilde{D}^T$ menggunakan definisi *norm* Forbenius pada teorema 2.24 persamaan (2.16).

$$\begin{aligned}\left\|A - \tilde{A}\right\|_F &= \sqrt{\sum_{i=1}^{12}\sum_{j=1}^{7}\left|a_{ij}\right|^2} \\ &= \sqrt{\sigma_4^2 + \sigma_5^2 + \sigma_6^2 + \sigma_7^2} \\ &\approx \sqrt{2.9099109} \\ &\approx 1.70585\end{aligned}$$

Jarak antara matriks awal dengan matriks aproksimasi dengan menggunakan *norm* Forbenius di atas memiliki nilai 1.70585. Hal ini menunjukkan bahwa dengan mengambil nilai $k = 3$, belum dapat memberikan hubungan sebenarnya dari kata

dengan dokumen. Oleh karena itu, pengambilan nilai k yang lebih besar akan memberikan hasil perkiraan lebih baik.

Akan tetapi, seperti yang dijelaskan sebelumnya, ekstraksi k topik meggunakan metode LSA akan memberikan hasil kurang tepat jika nilai k melebihi jumlah topik yang sebenarnya. Oleh karena itu penentuan nilai k membutuhkan suatu metode sehingga memberikan hasil yang lebih baik.

BAB 4
SIMULASI

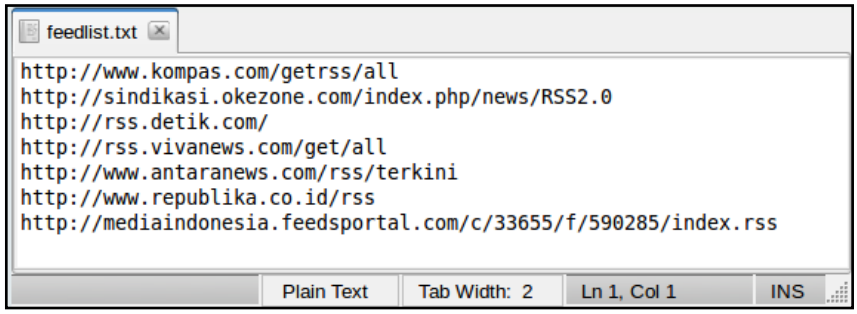
Bab ini akan menjelaskan hasil simulasi penggunaan LSA dengan SVD untuk mendapatkan topik utama harian dari portal berita Indonesia *online* selama satu bulan dengan keluaran selama lima hari yang diinterpretasikan dalam tugas akhir ini. Simulasi menggunakan bantuan program komputer dalam bahasa Python. Dengan spesifikasi prosesor: AMD Turion^{TM} X2 Dual-Core Mobile RM-74 (2CPUs), ~2.2 GHz, memori: 2048 MB RAM, sistem operasi: Ubuntu Ultimate Edition, kernel 2.6.2.2 , perangkat lunak : Python 2.7.2, modul tambahan: Numpy 1.6.1, Feedparser 5.1. Program dijalankan satu kali setiap harinya.

4.1 Akuisi Dokumen Berita

Proses pertama dari simulasi adalah menentukan dan mengumpulkan data yang akan digunakan. Simulasi pada tugas akhir ini menggunakan data yang berasal dari *file* RSS beberapa portal berita nasional. Portal berita *online* yang digunakan dalam penelitian ini berasal dari situs berita:

(a) Kompas.com
(b) Okezone.com
(c) Detik.com
(d) Vivanews.com
(e) Antaranews
(f) Republika *online*
(g) Media Indonesia

Alamat sumber *file* RSS dari seluruh portal berita *online* nasional yang digunakan di tuliskan dalam *file* ‘feedlist.txt’.



Gambar 4.1 *File* 'feedlist.txt'

Setelah penentuan sumber data, selanjutnya program 'kumpuldata.py' dijalankan untuk mendapatkan *file* RSS yang disediakan masing-masing portal berita. Algoritma pengumpulan data berupa *file* RSS yang dikerjakan oleh program digambarkan dalam bagan di bawah ini.

Algoritma pengumpulan data

Input : daftar alamat *feed* RSS.
Output : *file* RSS.

STEP 1 *set* feedlist = daftar alamat *feed* RSS
STEP 2 *for* feed *in* feedlist do *STEP 3-6*
STEP 3 *open* feed
STEP 4 *try* : *read* feed do STEP 5
STEP 5 *OUTPUT* (*write* feed *as* feed.rss)
('feed berhasil diunduh')
STEP 6 *except* : OUTPUT ('feed gagal diunduh')
STEP 7 *close* feed
STEP 8 *OUTPUT* ('Pengambilan selesai')
STOP.

Setiap waktu pengambilan, akan dilakukan percobaan untuk mengunduh *file* RSS untuk disimpan di dalam folder yang sesuai dengan hari pengambilan data. Pencatatan *file* terunduh maupun gagal terunduh disimpan dalam suatu *file* teks

bernama ‘logdata(tanggal pengambilan data).txt’. Gambar berikut menunjukkan contoh isi *file* ‘logdata5-1.txt’ yang mencatat keberhasilan dan kegagalan mengunduh *file* RSS pada tanggal 1 Mei 2012 pada pukul 13.55 WIB.

```
#--Log tanggal data5-1 2012

#-- Pada waktu : 13-55
Berhasil mengambil data http://www.kompas.com/getrss/all

Berhasil mengambil data http://sindikasi.okezone.com/index.php/news/RSS2.0

gagal mengambil data http://rss.detik.com/

Berhasil mengambil data http://rss.vivanews.com/get/all

Berhasil mengambil data http://www.antaranews.com/rss/terkini

Berhasil mengambil data http://www.republika.co.id/rss

Berhasil mengambil data http://mediaindonesia.feedsportal.com/c/33655/f/590285/index.rss
```

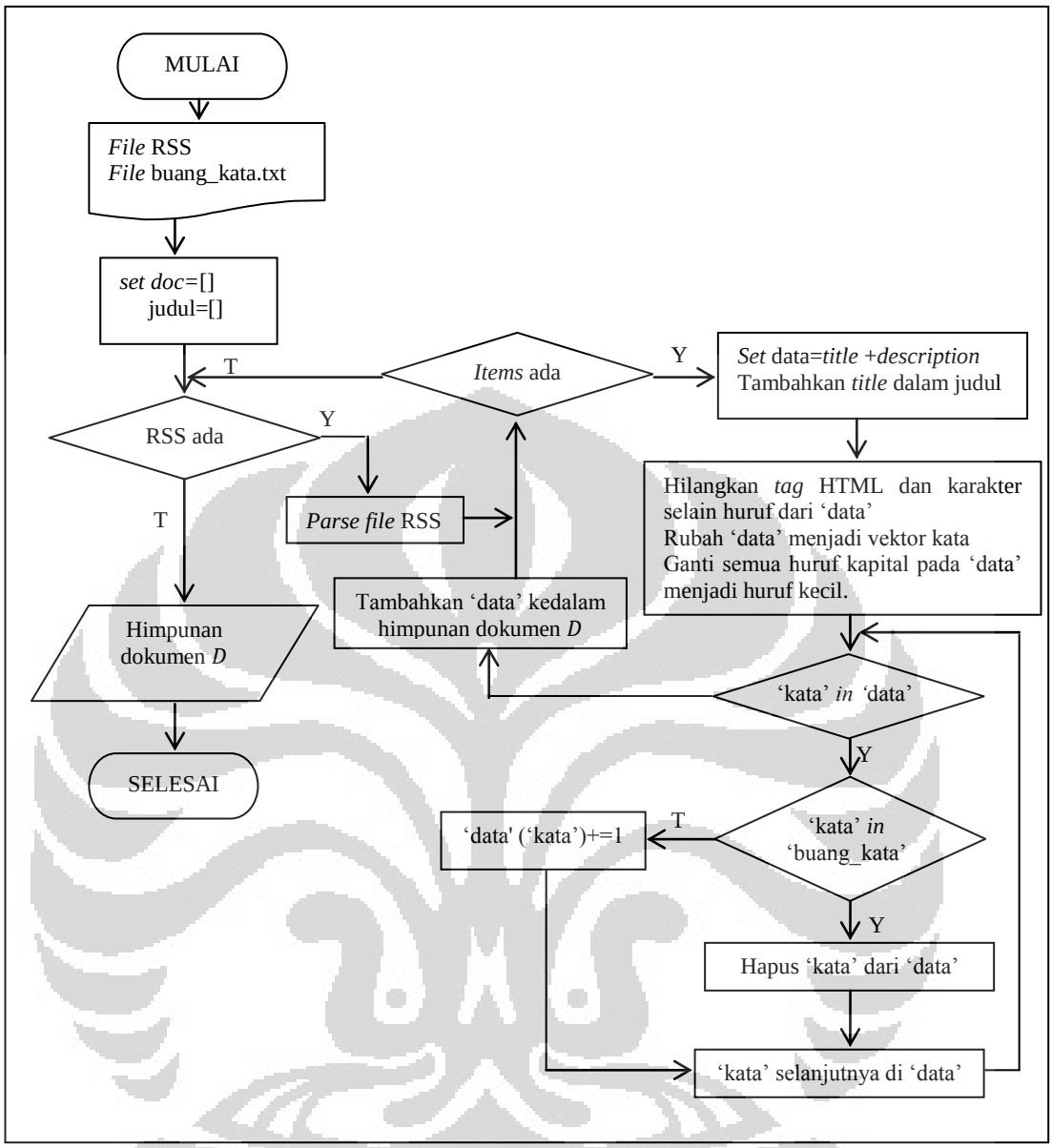
Gambar 4.2 Isi *file* ‘logdata5-1.txt’

4.2 Pernyiapan Dokumen Berita

4.2.1 Penyiapan Dokumen Berita

Dijelaskan pada subbab 3.3, sebelum disusun menjadi sebuah matriks kata-dokumen, dokumen-dokumen berita yang sudah terkumpul akan melalui beberapa tahap. Diawali dengan penghilangan *tag* HTML atau format, *tokenization, filtering* dan *weighting.* Proses penyiapan data ini digambarkan dalam sebuah *flowchart* pada gambar 4.3.

Kata-kata yang tidak diikutsertakan dalam pengolahan data tersimpan dalam *file* ‘buang_kata.txt’ terdapat di Lampiran 4. Hasil dari *flowchart* pada gambar 4.3 adalah himpunan dokumen *D* dimana setiap elemennya merupakan kata dan frekuensi kata dalam dokumen seperti contoh pada tabel 3.3 kolom keempat.



Gambar 4.3 *Flowchart* penyiapan data

4.2.2 Pembentukan Matriks kata-dokumen

Setelah mendapatkan himpunan vektor dokumen yang setiap elemennya merupakan kata dan frekuensi kata dalam dokumen, selanjutnya dibangun sebuah matriks kata-dokumen yang akan diproses SVD. Berikut ini algoritma pembentukan matriks kata-dokumen dari himpunan dokumen *D* pada proses sebelumnya.

Algoritma Pembentukan Matriks Kata-Dokumen

Input : himpunan vektor dokumen D.

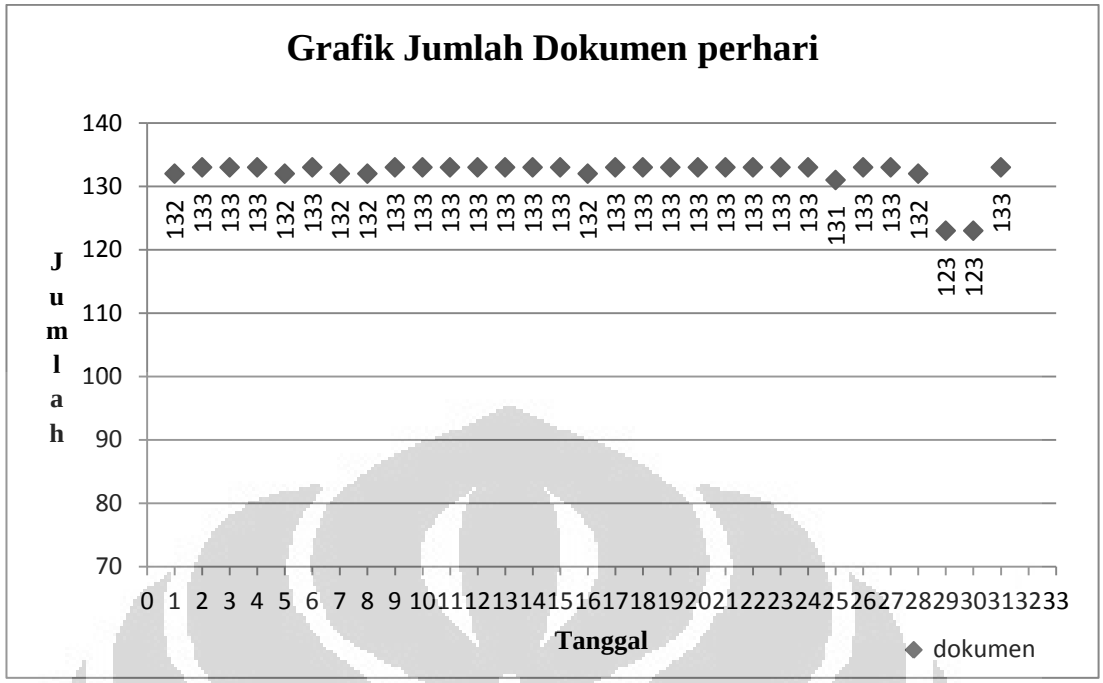
Output : matriks kata-dokumen, vektor kata dan vektor judul dokumen

STEP 1 *set* kata=[]
matriks=[]
n = *size*(D)
STEP 2 *for* 'dok' *in D do* ***STEP 3 -8***
STEP 3 *for* 'k' *in* 'dok' *do* ***STEP 4-8***
STEP 4 *if* k *not in* kata *do* ***STEP 5-8***
STEP 5 *kata.append*(k)
STEP 6 *for i=1 to n do* ***STEP 7***
STEP 7 matriks[k][i] = 0.
STEP 8 matriks[k][dok]+=1.
STEP 9 *Output* ('proses berhasil')
(matriks)

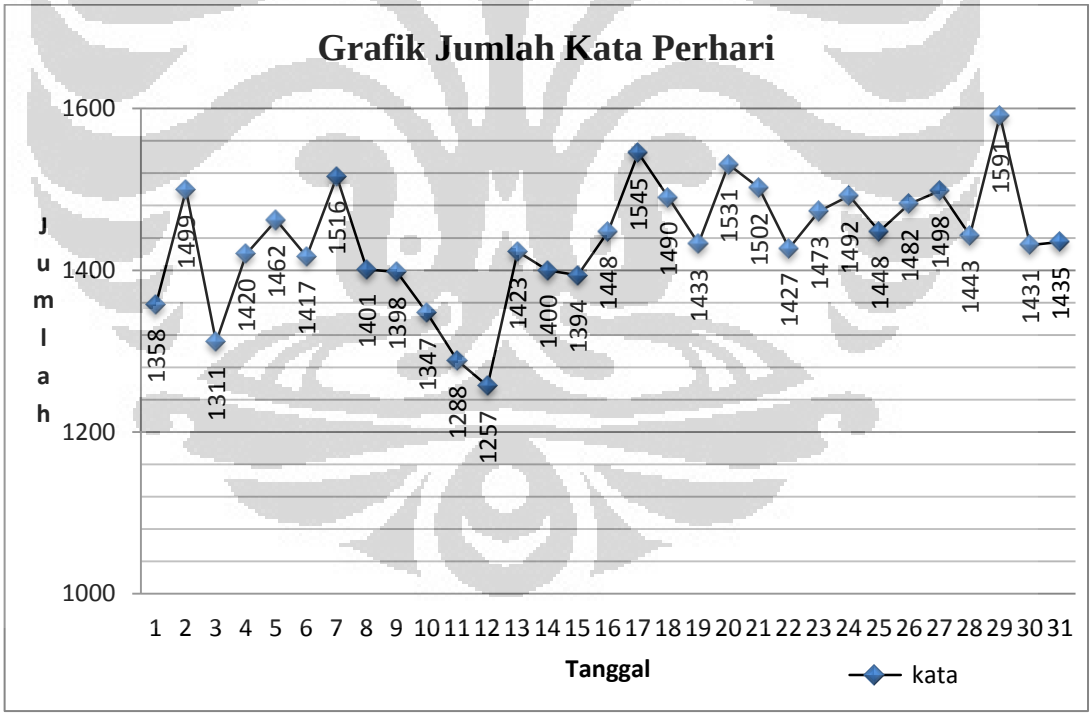
Berdasarkan teorema 2.23 yang menyebutkan bahwa faktorisasi SVD dapat dilakukan untuk sembarang matriks berukuran mxn, maka algoritma di atas tidak memerlukan syarat matriks yang dibentuk. Dokumen berita tidak mengandung seluruh kata dalam *dictionary* atau kamus, maka matriks yang dihasilkan memiliki banyak anggota bernilai 0.

Dengan menggunakan ukuran matriks kata-dokumen, didapat jumlah dokumen dan kata yang terambil setiap harinya. Grafik pada gambar 4.4 menggambarkan jumlah dokumen yang terambil dalam proses penyiapan selama bulan Mei 2012. Sedangkan gambar 4.5 menunjukkan grafik jumlah kata yang berbeda dalam proses ekstraksi topik dengan LSA selama bulan Mei 2012.

Gambar 4.4 dan gambar 4.5 menunjukkan ukuran matriks yang digunakan dalam percobaan selama bulan Mei 2012 memiliki jumlah kolom antara 123-133 dan jumlah baris antara 1257-1591.



Gambar 4.4 Grafik jumlah dokumen selama bulan Mei 2012

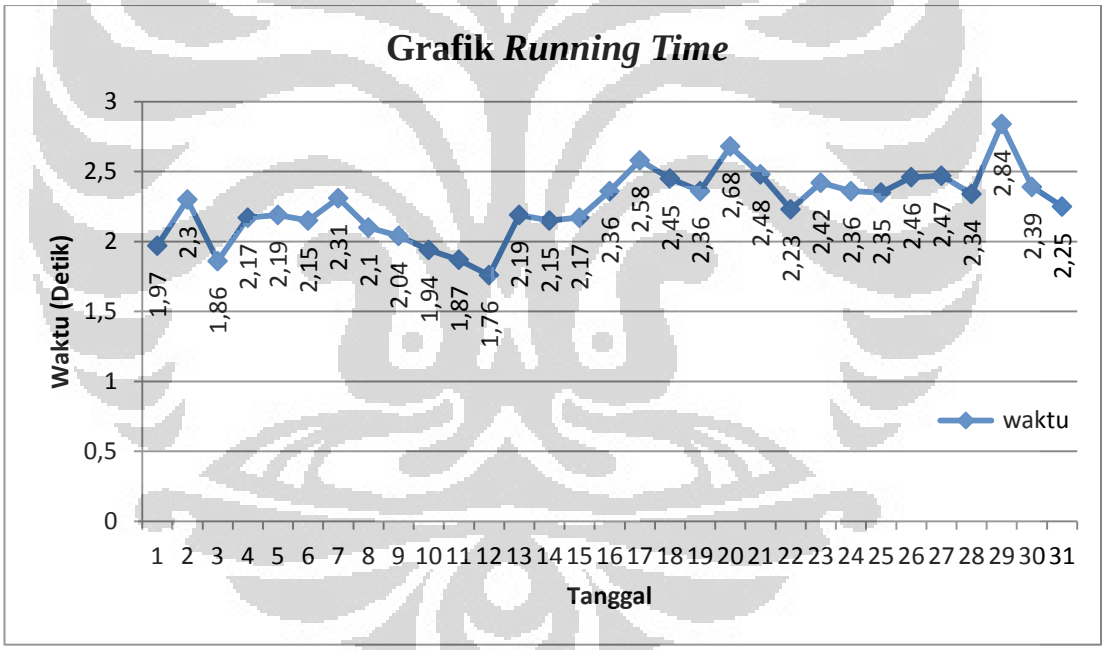


Gambar 4.5 Grafik jumlah kata berbeda yang dilibatkan selama bulan Mei 2012

4.3 Simulasi Ekstraksi Topik dengan Metode LSA

Tahapan selanjutnya dari metode LSA adalah faktorisasi SVD terhadap matriks kata-dokumen yang telah dibentuk pada tahap penyiapan data. Pada proses simulasi metode LSA selama bulan Mei 2012 dengan menggunakan program dan spesifikasi mesin yang dijelaskan pada bagian awal bab ini memberikan *running* time yang relatif kecil, sekitar 1,76-2,84 detik.

Grafik pada gambar 4.7 menunjukkan waktu atau *running time* yang dibutukan program untuk mengekstraksi topik selama bulan Mei 2012. *Running time* ini merupakan waktu yang dibutuhkan program untuk melakukan penyiapan data hingga menghasilan keluaran seperti pada gambar 4.6. Jumlah data yang diolah dijelaskan oleh gambar 4.4 dan gambar 4.5.



Gambar 4.6 *Running time* untuk ekstraksi topik utama

Tahap terakhir proses LSA adalah membuat keluaran dari program berupa *file* HTML yang berisikan :

(a) Tabel yang berisikan kata-kata yang tersekstraksi pada setiap topik beserta nilai hubungan kata dengan topik;

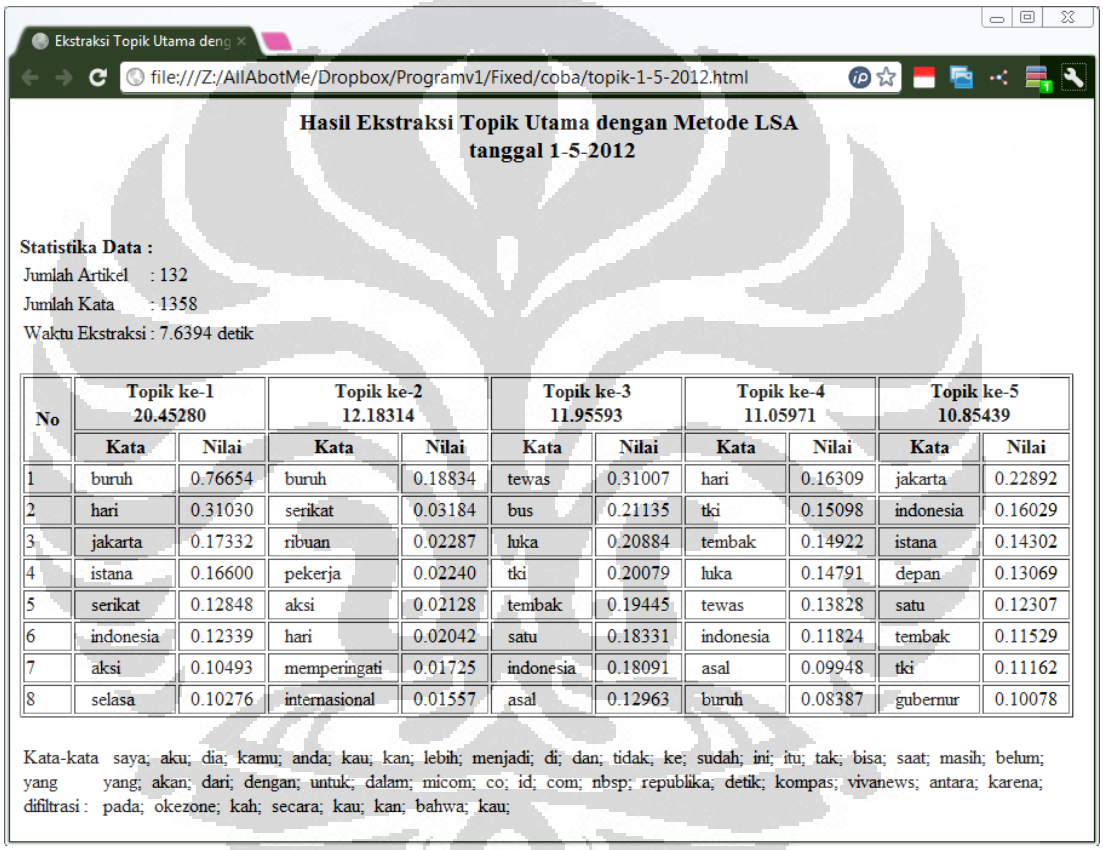
(b) Keterangan tanggal ekstraksi;

(c) Jumlah dokumen (artikel) dan kata-kata berbeda yang terlibat;

(d) *Running time* dari penyiapan data, pembentukan matriks kata-dokumen dan ekstraksi kata-kata;

(e) Kata-kata yang tidak diikutsertakan dalam proses (*filtrasi*).

Gambar 4.6 di bawah ini merupakan contoh *file* keluaran program yang sudah dikompilasi menggunakan *web browser* Chrome.



Ekstraksi Topik Utama deng ×

file:///Z:/AllAbotMe/Dropbox/Programv1/Fixed/coba/topik-1-5-2012.html

Hasil Ekstraksi Topik Utama dengan Metode LSA
tanggal 1-5-2012

Statistika Data :
Jumlah Artikel : 132
Jumlah Kata : 1358
Waktu Ekstraksi : 7.6394 detik

No	Topik ke-1 20.45280		Topik ke-2 12.18314		Topik ke-3 11.95593		Topik ke-4 11.05971		Topik ke-5 10.85439	
	Kata	Nilai	Kata	Nilai	Kata	Nilai	Kata	Nilai	Kata	Nilai
1	buruh	0.76654	buruh	0.18834	tewas	0.31007	hari	0.16309	jakarta	0.22892
2	hari	0.31030	serikat	0.03184	bus	0.21135	tki	0.15098	indonesia	0.16029
3	jakarta	0.17332	ribuan	0.02287	luka	0.20884	tembak	0.14922	istana	0.14302
4	istana	0.16600	pekerja	0.02240	tki	0.20079	luka	0.14791	depan	0.13069
5	serikat	0.12848	aksi	0.02128	tembak	0.19445	tewas	0.13828	satu	0.12307
6	indonesia	0.12339	hari	0.02042	satu	0.18331	indonesia	0.11824	tembak	0.11529
7	aksi	0.10493	memperingati	0.01725	indonesia	0.18091	asal	0.09948	tki	0.11162
8	selasa	0.10276	internasional	0.01557	asal	0.12963	buruh	0.08387	gubernur	0.10078

Kata-kata yang difiltrasi : saya; aku; dia; kamu; anda; kau; kan; lebih; menjadi; di; dan; tidak; ke; sudah; ini; itu; tak; bisa; saat; masih; belum; yang; akan; dari; dengan; untuk; dalam; micom; co; id; com; nbsp; republika; detik; kompas; vivanews; antara; karena; pada; okezone; kah; secara; kau; kan; bahwa; kau;

Gambar 4.7 *File* keluaran dari program

Sebagai contoh, tabel 4.1 sampai tabel 4.6 menunjukkan keluaran dari simulasi algoritma LSA untuk kumpulan dokumen berita pada tanggal 1-5 Mei 2012. Tabel tersebut berisikan kata-kata dan nilai yang menunjukkan hubungan kata dengan topik. Selanjutnya, untuk mengetahui kalimat topik, penulis mencoba untuk menginterpretasikan kata-kata menjadi suatu kalimat dan melakukan analisa.

Tabel 4.1 Keluaran metode LSA pada 1 Mei 2012 pukul 14.00 WIB

No	Topik ke-1 20.45280		Topik ke-2 12.18314		Topik ke-3 11.95593		Topik ke-4 11.05971		Topik ke-5 10.85439	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	buruh	0.76654	buruh	0.18834	kpk	0.31136	jalan	0.48074	jakarta	0.22892
2	hari	0.31030	serikat	0.03184	dpr	0.17555	merdeka	0.21266	indonesia	0.16029
3	jakarta	0.17332	ribuan	0.02287	riau	0.15879	massa	0.18517	istana	0.14302
4	istana	0.16600	pekerja	0.02240	ketua	0.15517	medan	0.17411	depan	0.13069
5	serikat	0.12848	aksi	0.02128	gubernur	0.14150	istana	0.16435	satu	0.12307
6	indonesia	0.12339	hari	0.02042	anis	0.13169	pendemo	0.16319	tembak	0.11529
7	aksi	0.10493	memperingati	0.01725	hari	0.11582	lintas	0.15854	tki	0.11162
8	selasa	0.10276	internasional	0.01557	rusli	0.10993	lalu	0.15854	gubernur	0.10078
	Hari buruh diperingati dengan aksi oleh serikat buruh Indonesia di Istana Negara.		Ribuan buruh dalam serikat pekerja memperingati hari buruh internasional dengan aksi.		Indikasi korupsi oleh KPK kepada anggota DPR, Anis dan gubernur Riau, Rusli.		Massa pendemo di jalan Medan Merdeka, depan Istana, mengganggu lalu lintas.		(sulit diinterpretasikan)	

Keterangan :

Keluaran metode LSA pada tanggal 1 Mei 2012 memperlihatkan topik tentang ‘peringatan aksi asosiasi pekerja pada peringatan hari buruh Internasional’ mendominasi. Topik ketiga mengekstrak kata yang merupakan nama yang terindikasi kasus korupsi. Ekstraksi topik kempat tetap memiliki kata-kata yang berkaitan dengan peringatan hari buruh. Sedangkan, dan kelima sulit untuk dapat diinterpretasikan karena terdiri dari kata-kata yang mendukung topik hari buruh (ke-1, 2 dan 4) serta gubernur dan penembakan TKI.

Tabel 4.2 Keluaran metode LSA pada 2 Mei 2012 pukul 14.00 WIB

No	Topik ke-1 12.97160		Topik ke-2 11.31134		Topik ke-3 10.98507		Topik ke-4 10.57992		Topik ke-5 9.66540	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	membimbing	-0.00017	kpk	0.32343	anggota	0.36429	indonesia	0.43786	ktp	0.49940
2	trofi	-0.00018	neneng	0.15354	dpr	0.22025	buruh	0.36324	januari	0.33294
3	sukses	-0.00018	nazaruddin	0.11170	hukuman	0.21697	seluruh	0.17406	bank	0.19428
4	real	-0.00018	korupsi	0.11127	ringan	0.19444	may	0.16681	tanah	0.18954
5	ranting	-0.00018	surat	0.07646	gorontalo	0.19285	day	0.16681	kendaraan	0.18171
6	meraih	-0.00018	sri	0.07286	brimob	0.19267	restoran	0.15258	simpan	0.17412
7	menyambut	-0.00018	minta	0.05841	komisi	0.15831	iswahyudi	0.15236	pengurusan	0.17384
8	mengungguli	-0.00018	kasus	0.05640	i	0.13323	rasa	0.14605	bermotor	0.17323
	Penyambutan Real Madrid yang sukses meraih trofi dengan mengungguli Liga Spanyol.		Nazaruddin meminta KPK mengirimkan surat kepada Neneng Sri Wahyuni untuk pengusutan kasus korupsi.		Anggota DPR dari komisi I menilai hukuman anggota brimob gorontalo.		Seluruh buruh Indonesia berunjuk rasa pada May Day.		Penggunaan e-KTP untuk pengurusan kendaraan bermotor dan simpanan di Bank mulai Januari	

Keterangan :

Hasil metode LSA pada tanggal 2 Mei ini menunjukkan topik-topik yang berbeda. Pada topik pertama, nilai dari kata-kata pendukung topik bernilai negatif, namun masih dapat diinterpretasikan menjadi sebuah kalimat topik.

Tabel 4.3 Keluaran metode LSA pada 3 Mei 2012 pukul 14.00 WIB

No	Topik ke-1 16.76140		Topik ke-2 15.86335		Topik ke-3 13.15896		Topik ke-4 10.89236		Topik ke-5 10.24442	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	endang	0.57679	kpk	0.40829	anis	0.21138	emas	0.48361	kpk	0.24445
2	rahayu	0.33568	ketua	0.31258	kpk	0.20804	perampok	0.34873	anas	0.24323
3	sedyaningsih	0.26444	anis	0.25050	matta	0.18165	cilacap	0.26563	korupsi	0.23902
4	kesehatan	0.19896	kasus	0.22577	dpr	0.11554	api	0.22644	komisi	0.17659
5	jenazah	0.19406	matta	0.21712	ppid	0.11262	toko	0.21385	pemberantasan	0.17343
6	kata	0.16366	partai	0.21253	wa	0.11120	lebeng	0.21385	demokrat	0.16371
7	sby	0.15913	jakarta	0.20661	ode	0.11120	matahari	0.20374	angelina	0.16006
8	mantan	0.15877	suap	0.16116	suap	0.10912	kg	0.18253	urbaningrum	0.12162
	Jenazah mantan menkses SBY, Endang Sri Rahayu.		Kasus suap wakil ketua Anis Matta oleh KPK.		Kasus suap yang diduga melibatkan anggota DPR Wa Ode dan Anis Matta terkait PPID.		Perampok toko emas di pasar Lebeng, Cilacap bersenjata api.		KPK, Komisi Pemberantasan korupsi menangani kasus Anas Urbaningrum, Angelina dari demokrat.	

Keterangan :

Ekstraksi topik pada tanggal 3 Mei ini didominasi dengan nama-nama tokoh nasional. Topik ke-1 berita tentang kematian menteri kesehatan, Endang Rahayu. Topik ke-2, ke-3 dan ke-5 memiliki kemiripan topik, yaitu kasus suap anggota DPR. Topik ke-4 terkait perampokan toko emas.

Tabel 4.4 Keluaran metode LSA pada 4 Mei 2012 pukul 14.00 WIB

No	Topik ke-1 18.18562		Topik ke-2 12.90919		Topik ke-3 12.01062		Topik ke-4 10.59956		Topik ke-5 10.34460	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	nba	0.33936	lt	0.00375	kasus	0.14865	ongen	0.34428	sq	0.23968
2	sportku	0.33285	iframe	0.00375	dhana	0.06025	munir	0.18147	penumpang	0.23644
3	lt	0.32892	nba	0.00349	ongen	0.05797	penumpang	0.18010	angie	0.19190
4	iframe	0.32892	sportku	0.00293	korupsi	0.05557	meninggal	0.16828	cengkareng	0.19093
5	www	0.21928	prediksi	0.00264	kejagung	0.04998	sq	0.16642	angelina	0.17935
6	width	0.21928	opening	0.00264	terkait	0.04940	keluarga	0.15545	a	0.16663
7	src	0.21928	finals	0.00264	kpk	0.04587	kematian	0.14672	ketua	0.15537
8	out	0.21928	www	0.00250	jakarta	0.04448	cengkareng	0.13279	indonesia	0.15037
	(tidak bisa diinterpretasikan)		Prediksi final NBA pada saat opening.		Kasus korupsi Dhana W di kejagung Jakarta.		Meninggalnya penumpang SQ di Cengakareng dan Ongen, saksi kasus Munir.		Kematian penumpang SQ di Cengkareng.	

Keterangan :

Topik pertama dan kedua memiliki kata-kata yang dalam *tag* HTML, hal ini menunjukkan berita yang disajikan dalam *file* RSS memiliki kesalahan pengkodean pada bagian berita. Namun, pada topik kedua, kata-kata dalam *tag* HTML berkurang, sehingga dapat diinterpretasikan. Selanjutnya untuk topik ketiga dan keempat tidak memiliki kesulitan untuk diinterpretasi. Topik kelima terlihat kata-kata yang terekstraksi sudah bercampur antara antara topik politik dan kematian penumpang.

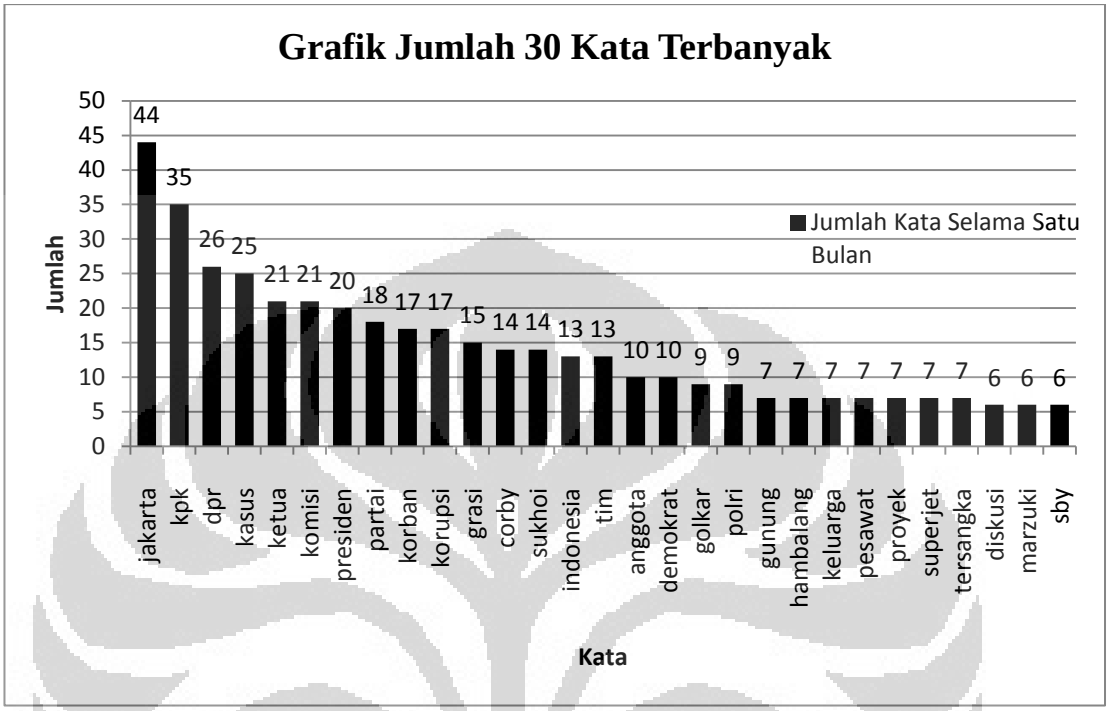
Tabel 4.5 Keluaran metode LSA pada 5 Mei 2012 pukul 14.00 WIB

No	Topik ke-1 12.93646		Topik ke-2 10.97600		Topik ke-3 9.71671		Topik ke-4 9.53970		Topik ke-5 9.31121	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	mc	-0.00015	jakarta	0.38477	kasus	0.35400	jakarta	0.11198	kasus	0.25911
2	jack	-0.00015	diskusi	0.25569	solo	0.18754	diskusi	0.08581	demokrasi	0.15685
3	tragis	-0.00015	salihara	0.22811	demokrasi	0.18574	salihara	0.07549	diskusi	0.13942
4	pingsan	-0.00015	jumat	0.20514	masuk	0.17585	buku	0.06206	kepala	0.12117
5	penyelaman	-0.00015	buku	0.16192	dhana	0.13317	manji	0.04787	masuk	0.12093
6	membentur	-0.00015	malam	0.14750	daerah	0.12689	pembubaran	0.04224	salihara	0.11914
7	kepalanya	-0.00015	selatan	0.14054	kepala	0.12686	komunitas	0.04123	keuangan	0.10503
8	francis	-0.00015	solo	0.13986	bui	0.12383	selatan	0.04019	bui	0.10456
	Penyelemana tragis Jack Francis MC yang pingsan akibat kapala yang terbentur boat		Diskusi buku di Salihara, Jakarta Selatan, Jumat malam.		(sulit diinterpretasikan)		Pembubaran diskusi buku Irshad Manji di komunitas Salihara.		(sulit diinterpretasikan)	

Keterangan :

Susunan kata-kata yang terekstraksi pada topik pertama memiliki nilai negatif yang berarti hubungan antara kata dan topik adalah negatif. Namun, masih dapat diinterpretasikan meski belum tentu merupakan topik utama sebenarnya seperti topik ke-1 pada tabel 4.2. Topik ke-3 dan ke-5 sulit diinterpretasikan karena mengekstraksi kata-kata yang bercampur.

Selanjutnya, dengan menghitung jumlah kata pendukung topik yang terekstraksi selama bulan Mei 2012, didapatkan 30 kata dengan jumlah terbanyak yang ditunjukkan dalam grafik di bawah ini.



Gambar 4.8 Grafik jumlah dari 30 kata terbanyak yang terekstraksi selama bulan Mei 2012.

Dengan menggunakan data yang terdapat di dalam grafik di atas, dapat dilakukan interpretasi topik utama selama satu bulan. Setidaknya, penulis dapat menginterpretasikan topik pada portal berita *online* berbahasa Indonesia dari sumber yang disebutkan di subbab 4.1 selama bulan Mei 2012 adalah:

(a) Kasus korupsi yang melibatkan anggota DPR pada beberapa proyek pemerintah,
(b) Pemberian grasi terhadap Corby,
(c) Tragedi jatuhnya pesawat Sukhoi superjet.

4.4 Perbandingan Hasil dengan Ekstraksi Manual

Setelah melakukan simulasi dan interpretasi pencarian topik utama berita harian dari portal berita *online*, pada bagian ini dilakukan perbandingan antara ektraksi topik menggunakan metode LSA dengan metode manual. Ekstraksi topik

secara manual dilakukan dengan membaca seluruh dokumen berita dan menyimpulkan beberapa topik menggunakan intuisi seorang relawan. Relawan merupakan seorang sarjana sastra.

Pencarian topik oleh relawan dilakukan secara independen, dimana relawan tidak mengetahui hasil keluaran yang didapat dari metode LSA. Relawan diberikan dokumen yang sama dengan data masukan metode LSA. Relawan diminta menentukan beberapa topik yang disusun berdasarkan dominasinya. Topik pertama merupakan topik paling dominan. Topik kedua adalah topik dominan selanjutnya dan seterusnya hingga topik kelima.

Perbandingan topik hasil interpretasi keluaran dengan metode LSA dan pencarian secara manual oleh relawan selama empat hari dituliskan dalam tabel 4.6 hingga tabel 4.9. Ekstraksi topik secara manual membutuhkan waktu hingga 3 jam setiap harinya.

Tabel 4.6 Perbandingan pada 1 Mei 2012 pukul 14.00 WIB

Topik ke	Interpretasi topik dengan metode LSA	Pencarian topik secara manual
1	Hari buruh diperingati dengan aksi oleh serikat buruh Indonesia di Istana Negara.	Demonstrasi warnai perayaan Hari Buruh
2	Ribuan buruh dalam serikat pekerja memperingati hari buruh internasional dengan aksi.	Kebakaran bus di Sumatra Barat tewaskan 13 penumpang
3	Indikasi korupsi oleh KPK kepada anggota DPR, Anis dan gubernur Riau, Rusli.	Dugaan keterlibatan Gubernur Riau dalam kasus suap PON
4	Massa pendemo di jalan Medan Merdeka, depan Istana, mengganggu lalu lintas.	Masalah Internal Golkar terkait pencalonan Ical
5	(sulit diinterpretasikan)	Pemanggilan Anis Matta oleh KPK terkait kasus suap DPID

Topik yang didapat antara metode LSA pada tanggal 1 Mei 2012 di Tabel 4.6 menunjukkan kemiripan topik pada topik kesatu, kedua dan keempat dengan topik kesatu pada pencarian secara manual. Topik ketiga pada LSA serupa dengan topik ketiga dan kelima pada metode manual.

Tabel 4.7 Perbandingan pada 2 Mei 2012 pukul 14.00 WIB

Topik ke	Interpretasi topik dengan metode LSA	Pencarian topik secara manual
1	Penyambutan Real Madrid yang sukses meraih trofi dengan mengungguli Liga Spanyol	Kondisi kesehatan Menkes semakin menurun
2	Nazaruddin meminta KPK mengirimkan surat kepada Neneng Sri Wahyuni untuk pengusutan kasus korupsi.	Nazaruddin surati KPK terkait koordinasi pemulangan Neneng
3	Anggota DPR dari komisi I menilai hukuman anggota brimob gorontalo.	Keterkaitan Anas dengan kasus Hambalang
4	Seluruh buruh Indonesia berunjuk rasa pada May Day.	KPK tawari Angie sebagai Justice Collaborator
5	Penggunaan e-KTP untuk pengurusan kendaraan bermotor dan simpanan di Bank mulai Januari	Hukuman anggota Brimob Gorontalo dinilai terlalu ringan

Tabel 4.7 menunjukkan topik terkait kondisi kesehatan menkes yang merupakan topik utama dari metode manual, tidak terekstraksi dengan metode LSA. Topik-topik yang memiliki kesamaan adalah topik ‘tentang pemulangan Neneng’ (topik kedua) dan ‘hukuman anggota brimob gorontalo’ (topik ketiga). Tiga topik lainnya tidak memiliki kesamaan.

Tabel 4.8 Perbandingan pada 3 Mei 2012 pukul 14.00 WIB

Topik ke	Interpretasi topik dengan metode LSA	Pencarian topik secara manual
1	Jenazah mantan menkses SBY, Endang Sri Rahayu	Pemakaman Menkes non-aktif Endang Rahayu
2	Kasus suap wakil ketua Anis Matta oleh KPK.	Dugaan keterlibatan Anis Matta dalam kasus PPID
3	Kasus suap yang diduga melibatkan anggota DPR Wa Ode dan Anis Matta terkait PPID.	SBY memberikan penghormatan terakhir untuk Endang Rahayu
4	Perampok toko emas di pasar Lebeng, Cilacap bersenjata api.	Meninggalnya saksi kunci kasus Munir
5	KPK, Komisi Pemberantasan korupsi menangani kasus Anas Urbaningrum, Angelina dari demokrat.	Kawanan perampok bersenjata api rampok toko emas di Cilacap

Tabel 4.8 menunjukkan topik yang memiliki kemiripan ada pada topik kesatu pada LSA dengan kesatu dan ketiga pada metode manual. Selanjutnya, topik kedua, ketiga dan kelima pada metode LSA serupa dengan topik ke kedua pada metode manual. Topik keempat pada metode LSA serupa dengan topik kelima pada metode manual. Sedangkan topik keempat metode manual tidak terekstraksi pada metode LSA.

Tabel 4.9 Perbandingan pada 4 Mei 2012 pukul 14.00 WIB

Topik ke	Interpretasi topik dengan metode LSA	Pencarian topik secara manual
1	(tidak bisa diinterpretasikan)	Kasus korupsi dan pencucian uang Dhana Widyatmika
2	Prediksi final NBA pada saat opening.	Bentrok warga dan laskar militant di Solo
3	Kasus korupsi Dhana W di kejagung Jakarta.	Kasus korupsi Angelina Sondakh
4	Meninggalnya penumpang SQ di Cengakareng dan Ongen, saksi kasus Munir.	Meninggalnya saksi kunci kasus Munir
5	Kematian penumpang SQ di Cengkareng.	Sutan Bhatoegana kunjungi Angie di tahanan

Tabel 4.9 menunjukkan dua topik yang memiliki kemiripan, yaitu topik ‘kasus korupsi Dhana Widyatmika’ dan ‘Meninggalnya saksi kunci kasus munir’.

Tabel 4.6 sampai tabel 4.9 menunjukkan beberapa kemiripan hasil dari interpretasi lima topik utama menggunakan metode LSA dengan metode manual. Kemiripan terbanyak terjadi pada ekstraksi tanggal 1 dan 3 Mei (4 topik, kemiripan 80%). Pada tanggal 2 dan 4 Mei hanya dua topik yang memiliki kemiripan atau kemiripan 40%. Dengan demikian, rata-rata kemiripan topik selama empat hari mencapai 60%.

Dengan kemiripan sebesar 60% pada perbandingan selama empat hari, maka metode LSA dapat membantu untuk mengekstraksi topik utama harian dari portal berita *online.*

BAB 5
KESIMPULAN DAN SARAN

Setelah melakukan percobaan implementasi metode LSA untuk mencari topik utama harian dari portal berita *online* berbahasa Indonesia, didapatkan hasil sebagai berikut :

(a) Program yang dibangun telah mampu mengumpulkan dokumen berita *online* berupa *file* RSS yang dapat digunakan kembali pada penelitian selanjutnya.

(b) Ekstraksi topik dengan menggunakan metode LSA yang terdiri dari kata-kata, sebagian besar sudah dapat diinterpretasikan menjadi kalimat topik secara manual.

(c) Perbandingan topik utama dengan metode LSA yang diinterpretasikan kedalam kalimat topik dengan ekatraksi topik secara manual selama empat hari menunjukkan kesamaan di berberapa topik setiap harinya.

Dengan hasil yang sudah didapat, maka dapat ditarik beberapa kesimpulan dan saran untuk perbaikan penelitian di masa depan.

5.1 Kesimpulan

Kesimpulan yang dapat dihasilkan dari percobaan penggunaan metode LSA untuk mengekstraksi topik utama adalah :

(a) Metode LSA dapat membantu ekstraksi topik dari kumpulan dokumen.

(b) Metode ini membutuhkan waktu yang relatif singkat dalam mencari topik utama dalam kumpulan dokumen seperti digambarkan pada gambar 4.7.

(c) Keluaran atau *output* bergantung pada kata-kata yang menyusun dokumen karena susunan kata berasal dari kata yang identik.

(d) Interpretasi hasil metode LSA membutuhkan pengetahuan yang cukup dari pengguna untuk memberikan tafsiran terbaik.

(e) Nilai negatif dalam hubungan kata dengan topik tetap memberikan interpretasi.

5.2 Saran

Untuk melengkapi metode LSA dengan faktorisasi SVD ini diperlukan perbaikan dalam penelitian lanjutan, diantaranya :

(a) Metode LSA pada penelitian ini masih membutuhkan interpretasi manual untuk mendapatkan kalimat topik. Akan tetapi, interpretasi manual bergantung pada pengetahuan dan intuisi pembaca. Dengan demikian, perlu untuk membangun algoritma lanjutan yang dapat menginterpretasi hasil secara otomatis.

(b) Program penelitian ini masih memberikan keluaran secara *offline* dan masih perlu mengunduh data berupa *file* RSS sehingga membutuhkan sarana penyimpanan yang besar. Implementasi secara *online* dapat dibangun pada masa depan untuk bisa memberikan informasi yang *realtime* dan tidak membutuhkan ruang penyimpanan yang besar karena dokumen dapat langsung di proses.

(c) Pada bagian akhir dari subbab 3.5 menunjukkan pengambilan nilai k-topik membutuhkan metode lain untuk memberikan hasil yang optimal.

DAFTAR ACUAN

Anton, Howard, Rorres, Chris. (2004). *Elementary Linear Algebra (9th Ed.).* United States: Wiley Book.

Burden, Richard L., Douglas. J Faires., dan Julet, Michael (ed.). (2011). *Numerical Analysis (9th Ed.).* Boston. USA : Brooks/Cole Cengage Learning.

Deerwester, Scott. dkk. (1990). Indexing by Latent Semantic Analysis. *Journal of American Society for Information Science* (1986-1998), 41(6):391-407.

Ghahramani, Zoubin. (2004) Unsupervised Learning. In Bousquet, O., von Luxburg, U. and Raetsch, G. *Advanced Lectures in Machine Learning*. Lecture Notes in Computer Science 3176, (pp: 72-112). Berlin: Springer-Verlag.

Golub, H, Gene. Van Loan, Charles, F. (1996). *Matrix Computations (3rd Ed.).* London : The Jhons Hopkins Press Ltd.

Honkela, Timo. Hyvärinen, Aapo. (2004). Linguistic Feature Extraction using Independent Component Analysis. *IEEE.* 279-284

Launder , T. K., Foltz, P. W., Laham, D. (1998). Introduction to Latent Sematic analysis. *Discourse Processes*, 25, 259-284

Murfi, Hendri. (2010). *Machine Learning for Text Indexing.* Disertasi. Berlin: Von der Fakultat IV -- Elektrotechnik und Informatik der Technischen Universitat.

Olney, Andrew. M. (2009). Generalizing Latent Semantic Analysis. *2009 IEEE International Conference on Semantic Computing.*

Pusat Bahasa Departemen Pendidikan Nasional (2008). *Kamus Bahasa Indonesia.* Jakarta: Departemen Pendidikan Nasional.

Segaran, Toby. (2007). *Programming Collective Intelligence.* Sebastopol, CA: O'Reilly Media, Inc.

LAMPIRAN

Lampiran 1

Listing program pengumpul data, *file* 'kumpuldata.py' .

```
#program untuk mengambil data dari berita online nasional
#created by : Ashari Nurhidayat

import time, os, re, urllib
import feedparser as fd

tgl=time.localtime()
folder='data'+str(tgl[1])+'-'+str(tgl[2])
fil=str(tgl[3])+'-'+str(tgl[4])+'.rss'
urllist=[line for line in file('feedlist.txt')]
#urllist=['antara.rss','detik.rss','kompas.rss','mi.rss']

#cek direktori dan masuk ke dalam direktori
if folder not in os.listdir('.'):
  os.mkdir(folder)
  logfile=file('log'+folder+'.txt','w')
  logfile.write('#--Log tanggal %s %d' %(folder,tgl[0]))
else :
  logfile=file('log'+folder+'.txt','a')
os.chdir(folder)
logfile.write('\n\n#-- Pada waktu : %d-%d' \
              %(tgl[3],tgl[4]))

for feed in urllist:
  try:
    f=urllib.urlopen(feed)
    g=f.read()
    h=fd.parse(g)
    out=h.feed.title.encode('utf8')+fil
    i=file(out,'w')
    i.write(str(g))
    i.close()
    f.close()
    logfile.write('\nBerhasil mengambil data %s' %(feed))
  except:
    logfile.write('\ngagal mengambil data %s' %(feed))
```

Lampiran 2

Listing program utama, *file* 'main.py'.

```
# Program utama untuk ekstraksi topik dari dokumen berita
melalui RSS
# Created by : Ashari Nurhidayat
# Versi : 1.0

import newsfeatures as nf
import os
from numpy import *
from time import clock,localtime

#pilih kata2 yang akan di hilangkan dari dalam folder
buang=[kata for kata in file('buang_kata.txt')]
buang_kata=[]
#hilangkan karakter \n,\r, ' ',',',';'
print 'Kata-kata yang dihilangkan : '
kar=['\n','\r',' ',',',';']
buang_kata=[]
for i in range(len(buang)):
  p=''
  for l in buang[i]:
    if l in kar :
      buang_kata.append(p)
      p=''
      continue
    p+=l
for i in buang_kata:
  if i=='': buang_kata.remove(i)
print buang_kata
t=clock()
#pilih folder
tgl=localtime()
folder='data'+str(tgl[1])+'-'+str(tgl[2])
os.chdir(folder)
#jumlah topik yang akan diekstrak
k=5
feedlist=os.listdir('.')
allw,artw,artt,errart=nf.getarticlewords(feedlist,buang_k
ata)
wordmatrix,wordvecs= nf.makematrix(allw,artw)
v=[]
v=matrix(wordmatrix)
v=transpose(v)
uk,sk,pk=nf.svddefrank(v,k)
os.chdir('..')
semua,uku,sing=nf.showfeatures(uk,sk,pk,wordvecs)
```

(Lanjutan)

```
waktu=clock()-t
er=' report berhasil '
#try:
nf.tulishtml(semua,buang_kata,tgl,waktu,uku,sing)
#except :
#  er=folder+ ' gagal menuliskan report'
#  print er
#news
#menuliskan dokumen yang gagal di kompilasi
f=file('log'+folder+'.txt','a')
f.write('\nPencarian  Topik  selesai\n-----%Dokumen  yang
bermasalah :\n')
if len(errart)>0:
  j=1
  for i in errart:
    f.write(str(j)+'. '+i+'\n')
    j+=1
else: f.write('Tidak ada dokumen bermasalah \n'+er)
print "Waktu dibutuhkan : "+str(waktu)
```

Lampiran 3

Listing modul *newsfeature*, *file* 'newsfeatures.py'.

```
# Modul Pelengkap untuk program ekstraksi topik berita
# dengan metode LSA
# Created by : Ashari Nurhidayat
# versi : 1.0

import feedparser
import re
from numpy import *
import numpy as np

# Menghapus gambar dan markup dari artikel
def getwords(html):
  # Penghapusan format HTML
  txt=re.compile(r'<[^>]+>').sub('',html)

  # Pemisahan kata berdasarkan karakter huruf
  words=re.compile(r'[^A-Z^a-z]+').split(txt)

  # konversi ke huruf kecil
  return [word.lower() for word in words if word!='']

# Mengambil kata dari artikel dan memisahkannya
def getarticlewords(feedlist,buang_kata):
  allwords={}
  errart=[]
  articlewords=[]
  articletitles=[]
  ec=0
  # Loop untuk setiap portal
  for feed in feedlist:
    f=feedparser.parse(feed)
    #loop untuk setiap artikel
    for e in f.entries:
      # mengabaikan artikel yang identik
      if e.title in articletitles: continue
      # ekstraksi kata
      try:
```

(Lanjutan)

```
        txt=e.title.encode('utf8')+'
'+e.description.encode('utf8')
      except:
        print 'artikel '+e.title.encode('utf8')+' : Tidak
lengkap'
        errart.append(e.title.encode('utf8'))
      words=getwords(txt)
      articlewords.append({} and '/n')
      articletitles.append(e.title)

      # filtering kata
      for word in words:
        if word in buang_kata : continue
        allwords.setdefault(word,0)
        allwords[word]+=1
        articlewords[ec].setdefault(word,0)
        articlewords[ec][word]+=1
      ec+=1
  return allwords,articlewords,articletitles,errart

# Fungsi pembentukan matriks kata-dokumen
def makematrix(allw,artlew):
  wordvec=[]
  # mengambil kata yang diduga sebagai topik
  for w,c in allw.items():
    wordvec.append(w)
    # membentuk matriks kata
  l1=[[(word in f and f[word] or 0) for word in wordvec]
for f in artlew]
  return l1, wordvec

# Fungsi pengambil kata-kata yang mendukung topik utama
def showfeatures(uk,sk,pk,wordvec):
  pd,wd=shape(uk) #uk(mxk)
  pc,wc=shape(pk) #pk(kxn)
  semua=[]
  sing=[]
  #for 1 to k-topic
  for i in range(wd):
    slist=[]
    n=[]
```

(Lanjutan)

```
    o=[]
    sing.append(sk[i][i])
    #search in topic-document matrix
    for j in range(pd):
      slist.append((uk[j,i],wordvec[j]))

    slist.sort()
    slist.reverse()
    #barisan kata2 pendukung topik
    #barisan nilai dari kata2 terhadap topik
    for s in slist[0:8]:
      n.append(s[1])
      if s[0] >(-0.0001) and s[0]<0.0001:
        if s[0]<0:
          o.append(-0.00000000001)
        else :
          o.append(0.000)
      else : o.append(s[0])
    m=[sum(s[0] for s in slist[0:8])]
    # menggabungkan dalam satu variabel
    semua.append(o)
    semua.append(n)

  si=[pd,wd,wc]
  return semua,si,sing

# Fungsi menghasilkan file keluaran berformat HTML
def tulishtml(semua,buang_kata,tgl,waktu,uku,sing):
  #menuliskan dalam file HTML
  tanggal='%d-%d-%d' %(tgl[2],tgl[1],tgl[0])
  bk=file('topik-'+tanggal+'.html','w')
  bk.write("""
  <html>
    <head>
      <title>Ekstraksi Topik Utama dengan Metode LSA pada
%s </title>
    </head>
  <body>"""%tanggal)
  bk.write("""<center>
    <h3>
```

(Lanjutan)

```
      <b>Hasil Ekstraksi Topik Utama dengan Metode
LSA</b>
    <br>tanggal %s</h3>
    </center><br><br>""" %tanggal)
  #Data Awal :
  bk.write("<b>Statistika Data :</b><br>")
  bk.write("""<table border=0>
    <tr>
      <td>Jumlah Artikel </td><td>: %d</td>
    </tr>
    <tr>
      <td>Jumlah Kata </td><td>: %d</td>
    </tr>
    <tr>
      <td>Waktu Ekstraksi </td><td>: %.4f detik
</td></table><br>"""%(int(uku[2]),int(uku[0]),waktu))
  bk.write("<table border=1>\n<tr><th rowspan=2>   No
  </th>")
  wd=uku[1]
  for i in range(wd):
    bk.write('<th colspan=2>Topik ke-%d<br> %.5f
</th>'%((i+1),sing[i]))
  bk.write('</tr><tr>')
  for i in range(wd):
    bk.write('<th> Kata </th><th> Nilai </th>')
  bk.write("</tr>\n")
  semua=transpose(semua)
  m,n=shape(semua)
  for i in range(m):
    bk.write('<tr><td> %2d </td>' %(i+1))
    for j in [0,2,4,6,8]:
      bk.write('<td>    '+semua[i][j+1]+' 
  </td>')
      type (semua[i][j])
      bk.write('<td align=right>   %.5f
 </td>'%(float(semua[i][j])))
    bk.write('</tr>')
  bk.write('</table><br> Kata-kata yang difiltrasi : ')
  for i in buang_kata:
    bk.write('  %s;'%i)
  bk.write('\n</body>\n</html>')
```

(Lanjutan)

```
  bk.close()

# fungsi defisiensi rank SVD
def svddefrank(v,k):
  u,s,p=np.linalg.svd(v)
  m=u.shape[0]
  n=p.shape[1]
  uk_s=s.size
  s=diag(s)
  s1=zeros((m,n))
  s1[:uk_s,:uk_s]=s
  U=u[:m,:k]
  S=s1[:k,:k]
  P=p[:k,:n]
  return U,S,P
```

Lampiran 4

File 'buang_kata.txt':

```
saya,aku,dia,kamu,anda,kau,kan,lebih,menjadi,di,dan,
tidak,ke,sudah,ini,itu,tak,bisa,saat,masih,belum,yang,
akan,dari,dengan,untuk,dalam,micom,co,id,com,nbsp,
republika,detik,kompas,vivanews,antara,karena,pada,
okezone,kah,secara,kau,kan,bahwa,kau
```

Lampiran 5

File 'feedlist.txt'.

```
http://www.kompas.com/getrss/all
http://sindikasi.okezone.com/index.php/news/RSS2.0
http://rss.detik.com/
http://rss.vivanews.com/get/all
http://www.antaranews.com/rss/terkini
http://www.republika.co.id/rss
http://mediaindonesia.feedsportal.com/c/33655/f/590285/in
dex.rss
```

Lampiran 6

Perintah pada program ‘crontab’.

```
# m h  dom mon dow   command
 59 23 * * * echo m4thui | sudo -S shutdown -h now
# pengumpulan rss setiap jam
 55          *          *          *          *          sh
/home/math/ayat/Dropbox/Programv1/programTestLab/Jadi/kum
pul.sh
 35          *          *          *          *          sh
/home/math/ayat/Dropbox/Programv1/programTestLab/Jadi/kum
pul.sh
#pencarian topik utama dalam 1 hari
 57         23          *          *          *          sh
/home/math/ayat/Dropbox/Programv1/programTestLab/Jadi/top
ik.sh
```

Lampiran 7

Listing bash file ‘kumpul.sh’.

```
cd /home/math/ayat/Dropbox/Programv1/programTestLab/Jadi
python kumpuldata.py
```

Lampiran 8

Listing bash file ‘topik.sh’.

```
cd /home/math/ayat/Dropbox/Programv1/programTestLab/Jadi
python main.py
```

Lampiran 9

Tabel-Tabel Hasil Ekstraksi Topik Utama Harian dengan Metode LSA Selama Bulan Mei 2012

Tanggal 1 Mei 2012

No	Topik ke-1 20.45280		Topik ke-2 12.18314		Topik ke-3 11.95593		Topik ke-4 11.05971		Topik ke-5 10.85439	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	buruh	0.76654	buruh	0.18834	kpk	0.31136	jalan	0.48074	jakarta	0.22892
2	hari	0.31030	serikat	0.03184	dpr	0.17555	merdeka	0.21266	indonesia	0.16029
3	jakarta	0.17332	ribuan	0.02287	riau	0.15879	massa	0.18517	istana	0.14302
4	istana	0.16600	pekerja	0.02240	ketua	0.15517	medan	0.17411	depan	0.13069
5	serikat	0.12848	aksi	0.02128	gubernur	0.14150	istana	0.16435	satu	0.12307
6	indonesia	0.12339	hari	0.02042	anis	0.13169	pendemo	0.16319	tembak	0.11529
7	aksi	0.10493	memperingati	0.01725	hari	0.11582	lintas	0.15854	tki	0.11162
8	selasa	0.10276	internasional	0.01557	rusli	0.10993	lalu	0.15854	gubernur	0.10078

Tanggal 2 Mei 2012

No	Topik ke-1 12.97160		Topik ke-2 11.31134		Topik ke-3 10.98507		Topik ke-4 10.57992		Topik ke-5 9.66540	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	membimbing	-0.00017	kpk	0.32343	anggota	0.36429	indonesia	0.43786	ktp	0.49940
2	trofi	-0.00018	neneng	0.15354	dpr	0.22025	buruh	0.36324	januari	0.33294
3	sukses	-0.00018	nazaruddin	0.11170	hukuman	0.21697	seluruh	0.17406	bank	0.19428
4	real	-0.00018	korupsi	0.11127	ringan	0.19444	may	0.16681	tanah	0.18954
5	ranting	-0.00018	surat	0.07646	gorontalo	0.19285	day	0.16681	kendaraan	0.18171
6	meraih	-0.00018	sri	0.07286	brimob	0.19267	restoran	0.15258	simpan	0.17412
7	menyambut	-0.00018	minta	0.05841	komisi	0.15831	iswahyudi	0.15236	pengurusan	0.17384
8	mengungguli	-0.00018	kasus	0.05640	i	0.13323	rasa	0.14605	bermotor	0.17323

Tanggal 3 Mei 2012

No	Topik ke-1 16.76140		Topik ke-2 15.86335		Topik ke-3 13.15896		Topik ke-4 10.89236		Topik ke-5 10.24442	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	endang	0.57679	kpk	0.40829	anis	0.21138	emas	0.48361	kpk	0.24445
2	rahayu	0.33568	ketua	0.31258	kpk	0.20804	perampok	0.34873	anas	0.24323
3	sedyaningsih	0.26444	anis	0.25050	matta	0.18165	cilacap	0.26563	korupsi	0.23902
4	kesehatan	0.19896	kasus	0.22577	dpr	0.11554	api	0.22644	komisi	0.17659
5	jenazah	0.19406	matta	0.21712	ppid	0.11262	toko	0.21385	pemberantasan	0.17343
6	kata	0.16366	partai	0.21253	wa	0.11120	lebeng	0.21385	demokrat	0.16371
7	sby	0.15913	jakarta	0.20661	ode	0.11120	matahari	0.20374	angelina	0.16006
8	mantan	0.15877	suap	0.16116	suap	0.10912	kg	0.18253	urbaningrum	0.12162

Tanggal 4 Mei 2012

No	Topik ke-1 18.18562		Topik ke-2 12.90919		Topik ke-3 12.01062		Topik ke-4 10.59956		Topik ke-5 10.34460	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	nba	0.33936	lt	0.00375	kasus	0.14865	ongen	0.34428	sq	0.23968
2	sportku	0.33285	iframe	0.00375	dhana	0.06025	munir	0.18147	penumpang	0.23644
3	lt	0.32892	nba	0.00349	ongen	0.05797	penumpang	0.18010	angie	0.19190
4	iframe	0.32892	sportku	0.00293	korupsi	0.05557	meninggal	0.16828	cengkareng	0.19093
5	www	0.21928	prediksi	0.00264	kejagung	0.04998	sq	0.16642	angelina	0.17935
6	width	0.21928	opening	0.00264	terkait	0.04940	keluarga	0.15545	a	0.16663
7	src	0.21928	finals	0.00264	kpk	0.04587	kematian	0.14672	ketua	0.15537
8	out	0.21928	www	0.00250	jakarta	0.04448	cengkareng	0.13279	indonesia	0.15037

Tanggal 5 Mei 2012

No	Topik ke-1 12.93646		Topik ke-2 10.97600		Topik ke-3 9.71671		Topik ke-4 9.53970		Topik ke-5 9.31121	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	mc	-0.00015	jakarta	0.38477	kasus	0.35400	jakarta	0.11198	kasus	0.25911
2	jack	-0.00015	diskusi	0.25569	solo	0.18754	diskusi	0.08581	demokrasi	0.15685
3	tragis	-0.00015	salihara	0.22811	demokrasi	0.18574	salihara	0.07549	diskusi	0.13942
4	pingsan	-0.00015	jumat	0.20514	masuk	0.17585	buku	0.06206	kepala	0.12117
5	penyelaman	-0.00015	buku	0.16192	dhana	0.13317	manji	0.04787	masuk	0.12093
6	membentur	-0.00015	malam	0.14750	daerah	0.12689	pembubaran	0.04224	salihara	0.11914
7	kepalanya	-0.00015	selatan	0.14054	kepala	0.12686	komunitas	0.04123	keuangan	0.10503
8	francis	-0.00015	solo	0.13986	bui	0.12383	selatan	0.04019	bui	0.10456

Tanggal 6 Mei 2012

No	Topik ke-1 16.63866		Topik ke-2 15.62320		Topik ke-3 11.83959		Topik ke-4 10.85161		Topik ke-5 10.55555	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	dpr	0.46667	jakarta	0.40938	senjata	0.09024	golkar	0.21684	kasus	0.25570
2	senjata	0.44206	irshad	0.39854	api	0.08917	partai	0.18449	kpk	0.21987
3	api	0.44117	manji	0.37868	anggota	0.06589	api	0.14190	jakarta	0.14688
4	anggota	0.38328	diskusi	0.34501	irshad	0.05197	senjata	0.13821	banggar	0.14245
5	komisi	0.15767	buku	0.23573	manji	0.04923	jk	0.09101	uang	0.13351
6	memiliki	0.10888	aji	0.15996	diskusi	0.03677	ical	0.08567	transaksi	0.12623
7	ketua	0.10086	malam	0.14984	buku	0.03106	bakrie	0.08567	ppatk	0.12303
8	iii	0.09880	pembubaran	0.13651	iii	0.02621	aburizal	0.08567	anggota	0.10812

Tanggal 7 Mei 2012

No	Topik ke-1 13.78770		Topik ke-2 12.23723		Topik ke-3 11.63757		Topik ke-4 10.68819		Topik ke-5 10.54775	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	tertimpa	-0.00000	senjata	0.43194	senjata	0.30653	partai	0.19167	partai	0.19823
2	terluka	-0.00000	api	0.37039	api	0.26206	demokrat	0.14707	indonesia	0.16899
3	putar	-0.00000	anggota	0.29093	anggota	0.18593	manji	0.11997	dua	0.14787
4	komidi	-0.00000	dpr	0.28320	dpr	0.16784	irshad	0.11997	golkar	0.12998
5	universitas	-0.00000	partai	0.18757	manji	0.11090	diskusi	0.10785	demokrat	0.11340
6	terima	-0.00000	polisi	0.15797	irshad	0.11090	presiden	0.10780	presiden	0.10835
7	kedokteran	-0.00000	demokrat	0.13895	polisi	0.10303	golkar	0.09530	siswa	0.10627
8	bengkulu	-0.00000	ketua	0.12980	diskusi	0.08941	wacana	0.09519	wacana	0.09657

Tanggal 8 Mei 2012

No	Topik ke-1 16.22897		Topik ke-2 12.09021		Topik ke-3 11.35797		Topik ke-4 10.05926		Topik ke-5 9.88032	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	kpk	0.47348	dpr	0.43937	dpr	0.18333	kasus	0.29142	marzuki	0.31391
2	kasus	0.28005	anggota	0.36472	kpk	0.14720	saksi	0.19310	alie	0.24422
3	dpr	0.27689	senjata	0.23921	anggota	0.13904	tersangka	0.14559	koruptor	0.22274
4	komisi	0.26178	senpi	0.21037	komisi	0.11966	anggota	0.13628	ui	0.19427
5	korupsi	0.23504	api	0.17864	neneng	0.09990	anggoro	0.10491	ketua	0.19249
6	jakarta	0.19558	dewan	0.16333	senpi	0.07516	periksa	0.10210	pernyataan	0.12579
7	anggota	0.19107	ketua	0.15350	wahyuni	0.07053	yulianis	0.10055	saksi	0.11992
8	ketua	0.16326	bk	0.13398	sri	0.07053	golkar	0.08974	alumni	0.11959

Tanggal 9 Mei 2012

No	Topik ke-1 15.18877		Topik ke-2 14.25426		Topik ke-3 13.70967		Topik ke-4 12.08419		Topik ke-5 11.58623	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	golkar	0.35669	golkar	0.23803	nunun	0.39009	nunun	0.42532	kpk	0.11055
2	partai	0.33451	partai	0.19470	kpk	0.21836	vonis	0.17916	kaban	0.05426
3	jakarta	0.27711	dewan	0.09559	kasus	0.20486	sidang	0.15033	ms	0.04597
4	kpk	0.24309	akbar	0.08698	vonis	0.15102	cek	0.14744	saksi	0.04538
5	nunun	0.20614	dpp	0.08640	tersangka	0.14479	adang	0.13215	diperiksa	0.04164
6	gorong	0.19178	ical	0.08469	korupsi	0.14232	golkar	0.13156	sebagai	0.04060
7	dewan	0.16712	pertimbangan	0.08254	jakarta	0.13514	pelawat	0.13093	mantan	0.04051
8	kasus	0.16603	rapat	0.08119	cek	0.12790	nurbaetie	0.12057	kasus	0.03531

Tanggal 10 Mei 2012

No	Topik ke-1 19.33166		Topik ke-2 12.62781		Topik ke-3 11.04095		Topik ke-4 10.50051		Topik ke-5 10.31924	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.60563	gunung	0.06131	kpk	0.32780	tiga	0.23665	pesawat	0.27476
2	pesawat	0.39962	pesawat	0.06025	suap	0.20693	jakarta	0.23367	salak	0.19429
3	superjet	0.24554	salak	0.05713	kasus	0.18684	malaysia	0.21546	gunung	0.17936
4	korban	0.21128	sukhoi	0.04807	menteri	0.17356	indonesia	0.20215	kecelakaan	0.11395
5	gunung	0.20236	tim	0.04785	dugaan	0.16543	wartawan	0.17196	penerbangan	0.09841
6	tim	0.19784	lokasi	0.03702	riau	0.16495	presiden	0.16139	jatuh	0.09797
7	salak	0.18987	sar	0.03434	keuangan	0.13734	golkar	0.14158	foto	0.09004
8	lokasi	0.15443	puing	0.02889	korupsi	0.12880	korban	0.12135	uu	0.07950

Tanggal 11 Mei 2012

No	Topik ke-1 20.68249		Topik ke-2 12.58922		Topik ke-3 11.30636		Topik ke-4 10.97373		Topik ke-5 9.97477	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.64313	sukhoi	0.20031	jakarta	0.28148	pesawat	0.25435	tim	0.19678
2	korban	0.32148	sar	0.18026	pesawat	0.26484	gunung	0.18237	sar	0.16851
3	pesawat	0.31665	tim	0.14774	gunung	0.22137	salak	0.15918	menkeu	0.15746
4	presiden	0.21267	gunung	0.12951	salak	0.19643	putin	0.09526	tersangka	0.15367
5	superjet	0.20633	salak	0.11913	eleven	0.13220	superjet	0.08820	kpk	0.14943
6	gunung	0.15288	tni	0.08029	salemba	0.13220	bogor	0.08018	agus	0.14889
7	salak	0.14289	jatuhnya	0.07782	aksi	0.12302	rusia	0.07318	bersaksi	0.12976
8	tim	0.12787	lokasi	0.07146	diduga	0.10719	presiden	0.06087	tni	0.12181

Tanggal 12 Mei 2012

No	Topik ke-1 21.77440		Topik ke-2 12.61513		Topik ke-3 12.06669		Topik ke-4 11.40540		Topik ke-5 10.69313	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.49640	korban	0.22990	jakarta	0.36257	jakarta	0.17700	marzuki	0.14864
2	korban	0.45295	sukhoi	0.22500	indonesia	0.21718	ketua	0.13436	partai	0.11600
3	jenazah	0.31625	evakuasi	0.17439	ketua	0.20656	partai	0.12105	indonesia	0.10672
4	pesawat	0.27690	tim	0.13972	marzuki	0.19964	marzuki	0.11827	dpr	0.09933
5	superjet	0.22007	pesawat	0.11978	dki	0.18947	dki	0.11172	nasdem	0.09337
6	kecelakaan	0.16387	salak	0.09111	ari	0.18809	indonesia	0.10786	jakarta	0.09253
7	halim	0.14864	gunung	0.08745	partai	0.17127	nasdem	0.08880	dewan	0.09141
8	polri	0.14706	proses	0.05732	dana	0.14562	dewan	0.07842	ketua	0.09029

Tanggal 13 Mei 2012

No	Topik ke-1 18.09934		Topik ke-2 11.18288		Topik ke-3 10.08187		Topik ke-4 9.45082		Topik ke-5 9.15633	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.55125	korban	0.09409	jenazah	0.54421	pengunggah	0.26252	data	0.27962
2	korban	0.43889	data	0.06168	kantong	0.32631	foto	0.25102	mortem	0.26001
3	tim	0.24844	mortem	0.05611	berisi	0.22850	pesawat	0.24690	jenazah	0.24019
4	superjet	0.23559	tim	0.03950	rs	0.16497	ys	0.17501	korban	0.23902
5	rusia	0.20157	post	0.03741	utuh	0.16425	meresahkan	0.17501	dvi	0.21287
6	pesawat	0.18284	ante	0.03741	polri	0.14955	berinisial	0.17501	post	0.17334
7	gunung	0.15586	dvi	0.03192	jakarta	0.09244	sukhoi	0.15633	ante	0.17334
8	jenazah	0.15431	superjet	0.02805	benda	0.08677	korban	0.14358	dna	0.12857

Tanggal 14 Mei 2012

No	Topik ke-1 15.05349		Topik ke-2 13.78869		Topik ke-3 12.21452		Topik ke-4 11.48360		Topik ke-5 10.71439	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	dpr	0.53565	sukhoi	0.47619	presiden	0.42331	merpati	0.46694	kabel	0.34790
2	anggota	0.30411	korban	0.26311	merpati	0.32947	setyopurnomo	0.20336	tiang	0.33400
3	sukhoi	0.26471	evakuasi	0.24845	jakarta	0.24385	rudy	0.20336	terkelupas	0.32950
4	bk	0.25273	gunung	0.23144	rapat	0.15261	pt	0.16000	dayat	0.27579
5	video	0.24394	rusia	0.22991	indonesia	0.14855	mogok	0.15736	kesetrum	0.25756
6	porno	0.20597	salak	0.22806	sebagai	0.13978	nusantara	0.15072	jalan	0.20817
7	kasus	0.15952	superjet	0.17055	golkar	0.13750	direktur	0.14965	lampu	0.19107
8	korban	0.15352	tim	0.15922	setyopurnomo	0.13640	airlines	0.14851	jakarta	0.16754

Tanggal 15 Mei 2012

No	Topik ke-1 14.23173		Topik ke-2 13.37858		Topik ke-3 12.67932		Topik ke-4 11.07523		Topik ke-5 10.34207	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.44135	sukhoi	0.45001	presiden	0.38335	presiden	0.27887	dpr	0.16991
2	jakarta	0.26614	superjet	0.17793	yudhoyono	0.24396	korea	0.16876	kasus	0.16597
3	konser	0.20136	gunung	0.14370	partai	0.20917	yong	0.12571	korea	0.16515
4	presiden	0.18466	korban	0.14092	demokrat	0.17481	nam	0.12571	jakarta	0.15950
5	lady	0.17712	tim	0.13858	korea	0.15611	kim	0.12571	presiden	0.15930
6	gaga	0.17712	lokasi	0.13121	capres	0.14177	konser	0.12437	kpk	0.13146
7	superjet	0.17348	pesawat	0.12860	susilo	0.12465	istana	0.12006	yong	0.12570
8	korban	0.14245	sar	0.12810	bambang	0.12465	utara	0.10770	nam	0.12570

Tanggal 16 Mei 2012

No	Topik ke-1 16.12141		Topik ke-2 13.24988		Topik ke-3 10.80723		Topik ke-4 10.29755		Topik ke-5 10.07092	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	hitam	0.44840	hitam	0.08344	partai	0.19058	presiden	0.31791	kotak	0.36187
2	sukhoi	0.41586	kotak	0.06001	pertemuan	0.14228	lady	0.21932	hitam	0.32928
3	kotak	0.36268	sukhoi	0.05370	prabowo	0.13751	konser	0.21932	jakarta	0.09332
4	ditemukan	0.21151	box	0.05016	yudhoyono	0.12828	gaga	0.21932	cvr	0.07637
5	knkt	0.20732	black	0.04954	presiden	0.12684	polda	0.21161	knkt	0.07328
6	black	0.19767	ditemukan	0.04114	pembina	0.12025	metro	0.21161	recorder	0.06139
7	box	0.19649	knkt	0.03626	dewan	0.12025	jaya	0.16877	kpk	0.06013
8	pesawat	0.19034	pesawat	0.03261	demokrat	0.11972	yong	0.11811	benda	0.05682

Tanggal 17 Mei 2012

No	Topik ke-1 12.92758		Topik ke-2 11.68958		Topik ke-3 10.97854		Topik ke-4 9.74429		Topik ke-5 9.56029	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	jaawkarta	0.35251	partai	0.29437	kpk	0.40938	kpk	0.23838	box	0.34010
2	sukhoi	0.28308	artis	0.24205	rekening	0.22512	box	0.18802	black	0.33443
3	kpk	0.27664	pks	0.20066	polri	0.18220	black	0.18614	tahun	0.20119
4	pesawat	0.19466	kpk	0.17577	kasus	0.17432	rekening	0.15390	satu	0.17754
5	partai	0.17985	jakarta	0.13378	korupsi	0.15026	polri	0.14164	kapal	0.16603
6	korban	0.14886	calon	0.11028	gendut	0.11265	artis	0.12144	as	0.15071
7	komisi	0.13998	merekrut	0.10510	pemberantasan	0.10298	kasus	0.11235	knkt	0.13998
8	rekening	0.13875	sejahtera	0.09827	komisi	0.08381	pks	0.10087	california	0.12452

Tanggal 18 Mei 2012

No	Topik ke-1 15.86615		Topik ke-2 11.98744		Topik ke-3 10.55239		Topik ke-4 10.23833		Topik ke-5 9.63033	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	empulse	0.00000	partai	0.46851	tim	0.54270	partai	0.27731	hakim	0.54621
2	brammo	0.00000	jakarta	0.29773	u	0.31767	demokrat	0.21492	ma	0.40122
3	top	0.00000	demokrat	0.29685	ngawi	0.21178	politik	0.17267	timur	0.27866
4	speed	0.00000	politik	0.22876	nasional	0.18614	kemanusiaan	0.11370	ribu	0.27209
5	ramah	0.00000	artis	0.15807	pemain	0.16302	anas	0.10587	ketua	0.16406
6	r	0.00000	calon	0.15469	persinga	0.16128	kegiatan	0.10315	sekitar	0.14124
7	motor	0.00000	anas	0.14799	belakang	0.16128	asuransi	0.07986	awasi	0.13818
8	meluncurkan	0.00000	presiden	0.14300	pelatih	0.15947	hal	0.07962	agung	0.13590

Tanggal 19 Mei 2012

No	Topik ke-1 15.78705		Topik ke-2 14.87198		Topik ke-3 13.44269		Topik ke-4 10.56988		Topik ke-5 10.18515	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	pecat	-0.00000	presiden	0.46285	gubernur	0.13367	tim	0.18100	dahlan	0.23935
2	simone	-0.00000	timor	0.36191	dpt	0.10905	pemilih	0.14343	mobil	0.20057
3	monaco	-0.00000	leste	0.36191	pemilih	0.08093	dpt	0.11446	tim	0.18687
4	marco	-0.00000	yudhoyono	0.25005	tetap	0.07296	data	0.11168	puncak	0.16359
5	unit	-0.00000	sby	0.20586	daftar	0.07229	timor	0.10544	tol	0.15366
6	sedikitnya	-0.00000	bambang	0.12947	penetapan	0.06230	leste	0.10544	data	0.15043
7	pemadam	-0.00000	susilo	0.12723	calon	0.05896	fdr	0.09996	novel	0.13680
8	dikerahkan	-0.00000	bertemu	0.12531	presiden	0.04353	kpu	0.09527	petugas	0.13660

Tanggal 20 Mei 2012

No	Topik ke-1 15.44439		Topik ke-2 11.65689		Topik ke-3 9.84222		Topik ke-4 9.27107		Topik ke-5 8.77644	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	tentang	0.00000	korban	0.08568	presiden	0.33851	piala	0.28881	kasus	0.31345
2	lalu	0.00000	sukhoi	0.08094	indonesia	0.18828	thomas	0.27372	century	0.22255
3	kisah	0.00000	dvi	0.02701	portugal	0.17221	matahari	0.22141	kpk	0.22115
4	iskan	0.00000	identifikasi	0.02600	bertemu	0.10714	hasil	0.21683	polisi	0.20228
5	diluncurkan	0.00000	keluarga	0.02401	yudhoyono	0.08232	pertama	0.19949	presiden	0.16845
6	bumn	0.00000	jasad	0.02359	pertama	0.07516	wuhan	0.19076	kejaksaan	0.15298
7	bercerita	0.00000	superjet	0.02176	berkunjung	0.07269	hari	0.18828	bank	0.15072
8	sepatu	0.00000	pesawat	0.02121	dijadwalkan	0.07050	indonesia	0.18418	belanda	0.14935

Tanggal 21 Mei 2012

No	Topik ke-1 16.61845		Topik ke-2 13.31643		Topik ke-3 10.70888		Topik ke-4 10.65044		Topik ke-5 9.70933	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.56509	sukhoi	0.05737	kpk	0.27927	fdr	0.38479	langkah	0.13431
2	korban	0.48369	korban	0.04654	pt	0.15212	pencarian	0.28584	komisi	0.13164
3	keluarga	0.24194	keluarga	0.02973	kasus	0.10334	gunung	0.23196	iii	0.12713
4	jasad	0.18904	jasad	0.02843	periksa	0.10061	salak	0.18780	dpr	0.11965
5	tim	0.18507	tim	0.02131	dugaan	0.09317	knkt	0.17489	agusrin	0.11827
6	gunung	0.16205	melihat	0.02005	terkait	0.08690	data	0.15593	tepat	0.08580
7	salak	0.15188	dvi	0.01807	mantan	0.07964	tim	0.15458	bengkulu	0.08062
8	superjet	0.14059	polri	0.01744	idris	0.07928	recorder	0.13337	fdr	0.07886

Tanggal 22 Mei 2012

No	Topik ke-1 14.61047		Topik ke-2 13.51997		Topik ke-3 11.99783		Topik ke-4 11.30507		Topik ke-5 10.88669	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	tahan	0.00000	sukhoi	0.34835	lady	0.44437	kpk	0.34468	kpk	0.33886
2	polres	0.00000	korban	0.33347	gaga	0.44437	jakarta	0.27793	gerhana	0.18458
3	membawa	0.00000	keluarga	0.30504	konser	0.42251	gerhana	0.19607	anggota	0.14164
4	kediri	0.00000	jenazah	0.16931	promotor	0.15598	sebagai	0.18907	panggilan	0.14057
5	kedapatan	0.00000	melihat	0.15712	ingin	0.15517	panggilan	0.14683	korupsi	0.13797
6	ganja	0.00000	jasad	0.13304	digelar	0.12809	menteri	0.14200	komisi	0.13521
7	dosen	0.00000	superjet	0.11304	daddy	0.10250	korupsi	0.13795	sianipar	0.12364
8	daun	0.00000	polri	0.10838	big	0.10250	sianipar	0.12945	dpr	0.11581

Tanggal 23 Mei 2012

No	Topik ke-1 15.20895		Topik ke-2 12.88900		Topik ke-3 12.18225		Topik ke-4 11.87534		Topik ke-5 11.42603	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	video	-0.00011	demokrat	0.24018	corby	0.38683	wa	0.28250	lady	0.27323
2	halaman	-0.00011	fraksi	0.22925	grasi	0.33637	ode	0.28250	gaga	0.27323
3	disemayamkan	-0.00011	partai	0.20752	ketua	0.19191	kasus	0.27489	konser	0.25167
4	keroyok	-0.00012	wa	0.18115	presiden	0.18408	corby	0.19834	dubes	0.17526
5	ulah	-0.00012	ode	0.18115	australia	0.18025	tersangka	0.18369	as	0.15593
6	nekat	-0.00012	ketua	0.17480	kepada	0.14578	grasi	0.17480	tim	0.12853
7	mengeroyok	-0.00012	nurhayati	0.15744	schapelle	0.12339	kpk	0.15708	polri	0.12826
8	membakar	-0.00012	kpk	0.14528	dpr	0.11463	dugaan	0.11892	fraksi	0.12142

Tanggal 24 Mei 2012

No	Topik ke-1 14.92291		Topik ke-2 12.26847		Topik ke-3 10.53421		Topik ke-4 10.01648		Topik ke-5 9.59578	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	grasi	0.42333	kpk	0.50870	kpk	0.30411	konser	0.27290	konser	0.24989
2	corby	0.39799	kasus	0.23783	andi	0.13154	lady	0.25032	lady	0.22170
3	australia	0.25643	andi	0.23331	hambalang	0.10874	gaga	0.25032	gaga	0.22170
4	pemberian	0.22288	korupsi	0.20494	grasi	0.09651	kpk	0.17536	ketua	0.11665
5	kasus	0.20720	hambalang	0.18128	corby	0.08724	pihak	0.14028	kasus	0.10994
6	jakarta	0.20461	jakarta	0.15882	olahraga	0.07697	indonesia	0.12847	wakil	0.08802
7	tahun	0.18972	menteri	0.14063	pemberantasan	0.07154	ketua	0.09818	polri	0.08537
8	schapelle	0.17860	komisi	0.13369	menteri	0.06964	polri	0.08519	izin	0.07752

Tanggal 25 Mei 2012

No	Topik ke-1 15.49766		Topik ke-2 13.18951		Topik ke-3 11.81443		Topik ke-4 10.44026		Topik ke-5 10.05721	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	propam	-0.00010	demokrat	0.25330	grasi	0.24173	un	0.24161	mobil	0.22654
2	melaporkan	-0.00010	ternate	0.20947	corby	0.19470	lulus	0.19890	tni	0.20688
3	kedua	-0.00010	anas	0.20476	presiden	0.13495	arus	0.17744	un	0.18868
4	istri	-0.00010	partai	0.20137	terpidana	0.10440	dua	0.17311	sipil	0.16154
5	ban	-0.00010	maluku	0.15633	yudhoyono	0.10202	sma	0.16472	lulus	0.15363
6	anaknya	-0.00010	utara	0.14156	schapelle	0.09065	jakarta	0.16403	i	0.13097
7	ain	-0.00010	urbaningrum	0.13729	narkoba	0.08907	nasional	0.15931	sma	0.12897
8	memperkuat	-0.00012	ketua	0.13514	ironis	0.08221	tewas	0.14357	fortuner	0.12500

Tanggal 26 Mei 2012

No	Topik ke-1 14.09432		Topik ke-2 12.00958		Topik ke-3 11.32468		Topik ke-4 10.45199		Topik ke-5 10.33579	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	tau	0.00000	siswa	0.43325	siswa	0.29200	buyung	0.36867	korupsi	0.16645
2	sini	0.00000	kelulusan	0.26428	lulus	0.18848	sby	0.34245	kpk	0.15316
3	setu	0.00000	lulus	0.25863	grasi	0.18099	adnan	0.20530	ketua	0.13700
4	rumahnye	0.00000	jakarta	0.21059	corby	0.16843	nasihat	0.17781	partai	0.12402
5	nggak	0.00000	ujian	0.18043	kelulusan	0.13919	buku	0.17178	demokrat	0.11248
6	manggung	0.00000	nasional	0.17339	snmptn	0.10591	mantan	0.16150	jakarta	0.08749
7	mana	0.00000	sma	0.17132	ujian	0.09865	anggota	0.15914	dugaan	0.06449
8	lagi	0.00000	sekolah	0.15471	sma	0.09471	nasution	0.13714	umum	0.06383

Tanggal 27 Mei 2012

No	Topik ke-1 16.42481		Topik ke-2 13.47809		Topik ke-3 12.33305		Topik ke-4 11.35233		Topik ke-5 10.10724	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	turun	0.00000	mkgr	0.22649	demokrat	0.33959	corby	0.21270	sby	0.22982
2	ramaikan	0.00000	ketua	0.22560	pd	0.23993	grasi	0.21036	buyung	0.14813
3	yogyakarta	0.00000	dewan	0.20057	partai	0.21765	pd	0.12574	pd	0.14599
4	www	0.00000	alex	0.16741	kader	0.12233	ketua	0.10011	adnan	0.11885
5	wajah	0.00000	golkar	0.16188	anas	0.11930	dewan	0.09684	yudhoyono	0.10915
6	viva	0.00000	kehormatan	0.14580	utara	0.10359	pemberian	0.09134	presiden	0.10896
7	memperkenalkan	0.00000	partai	0.13275	maluku	0.10062	sby	0.08100	ani	0.09613
8	funbike	0.00000	noerdin	0.12928	penyerangan	0.09984	mkgr	0.06765	anggota	0.09136

Tanggal 28 Mei 2012

No	Topik ke-1 14.44588		Topik ke-2 12.02655		Topik ke-3 10.79601		Topik ke-4 10.25553		Topik ke-5 9.95127	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	tembak	0.00000	hambalang	0.40946	hambalang	0.28381	jakarta	0.24660	corby	0.19756
2	sambut	0.00000	proyek	0.27549	sukhoi	0.22978	ii	0.23857	grasi	0.18021
3	papua	0.00000	kpk	0.18377	proyek	0.19890	atc	0.21122	dpr	0.17417
4	pangdam	0.00000	bangunan	0.14282	ii	0.15543	pura	0.14452	hambalang	0.14057
5	meninggalkan	0.00000	kasus	0.14075	atc	0.14073	angkasa	0.14452	marzuki	0.11789
6	kapolda	0.00000	olahraga	0.13062	kaki	0.11146	polri	0.14387	presiden	0.10500
7	cendrawasih	0.00000	lokasi	0.11057	turun	0.10144	gaga	0.13781	ketua	0.10453
8	baku	0.00000	tersebut	0.10720	pura	0.10089	kaki	0.13248	interpelasi	0.10261

Tanggal 29 Mei 2012

No	Topik ke-1 13.44943		Topik ke-2 11.01124		Topik ke-3 10.15305		Topik ke-4 9.27433		Topik ke-5 9.13980	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	dpr	0.38731	dpr	0.38467	kpk	0.29890	jakarta	0.28936	kasus	0.28214
2	komisi	0.35201	ri	0.17637	hambalang	0.23600	sukhoi	0.17419	korupsi	0.18037
3	jakarta	0.25735	anggota	0.16982	komisi	0.22427	izin	0.16375	ri	0.15941
4	anggota	0.20010	x	0.16327	korupsi	0.20327	ri	0.11797	angie	0.13312
5	kasus	0.19598	komisi	0.13549	angie	0.18633	rusia	0.10442	tersangka	0.11744
6	x	0.18946	bk	0.11190	proyek	0.18259	telah	0.09758	dugaan	0.11729
7	hambalang	0.18595	m	0.10592	pemberantasan	0.14249	masalah	0.09396	investasi	0.10396
8	ketua	0.17131	hambalang	0.10537	pembangunan	0.14205	jokowi	0.09396	kembali	0.10094

Tanggal 30 Mei 2012

No	Topik ke-1 14.72668		Topik ke-2 12.59949		Topik ke-3 11.57886		Topik ke-4 10.46342		Topik ke-5 10.06606	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	wibowo	0.00000	tni	0.45375	century	0.17587	corby	0.36281	dpr	0.29175
2	sunu	0.00000	nomor	0.26765	dpr	0.17073	grasi	0.33516	ketua	0.28269
3	secuil	0.00000	pelat	0.24615	kpk	0.11298	presiden	0.17460	komisi	0.23064
4	pondok	0.00000	jalan	0.15557	komisi	0.10684	indonesia	0.16244	ix	0.17632
5	menikmati	0.00000	palsu	0.15419	bank	0.08217	deal	0.16085	hambalang	0.14128
6	harmonisasi	0.00000	plat	0.14709	ketua	0.07578	dpr	0.16061	wakil	0.13973
7	galeri	0.00000	mobil	0.14514	rapat	0.07133	schapelle	0.15626	proyek	0.13147
8	damianus	0.00000	pinggir	0.11698	timwas	0.07043	pemberian	0.14441	anggota	0.12493

Tanggal 31 Mei 2012

No	Topik ke-1 14.17738		Topik ke-2 13.12579		Topik ke-3 11.23787		Topik ke-4 10.57859		Topik ke-5 9.52410	
	Kata	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**	**Kata**	**Nilai**
1	sukhoi	0.52160	sukhoi	0.15880	tni	0.60428	kpk	0.23393	pohon	0.26383
2	fdr	0.49164	fdr	0.15809	panglima	0.35333	dpr	0.17526	jakarta	0.25287
3	ditemukan	0.23753	ditemukan	0.08150	wartawan	0.23502	angelina	0.15979	tumbang	0.23217
4	superjet	0.17972	superjet	0.05665	padang	0.23456	proyek	0.15015	hari	0.16660
5	salak	0.17249	recorder	0.05415	insiden	0.12924	tni	0.13517	kretek	0.15388
6	gunung	0.17249	data	0.05415	tempat	0.12024	komisi	0.10996	komunitas	0.15388
7	recorder	0.15866	salak	0.05388	asusila	0.11891	korupsi	0.10170	juga	0.13580
8	data	0.15866	gunung	0.05388	praktik	0.11891	hambalang	0.10096	medan	0.13516