



**UNIVERSITAS INDONESIA**

**PENGELOMPOKAN DOKUMEN BAHASA INDONESIA  
DENGAN TEKNIK REDUKSI DIMENSI NONNEGATIVE  
MATRIX FACTORIZATION DAN RANDOM PROJECTION**

**SKRIPSI**

**Suryanto Ang  
1205000886**

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS INDONESIA  
DEPOK  
2009**



**UNIVERSITAS INDONESIA**

**PENGELOMPOKAN DOKUMEN BAHASA INDONESIA  
DENGAN TEKNIK REDUKSI DIMENSI NONNEGATIVE  
MATRIX FACTORIZATION DAN RANDOM PROJECTION**

**SKRIPSI**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana**

**Suryanto Ang  
1205000886**

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS INDONESIA  
DEPOK  
2009**

## **HALAMAN PERNYATAAN ORISINALITAS**

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

**Nama : Suryanto Ang**

**NPM : 1205000886**

**Tanda Tangan :**

**Tanggal :**



## HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :  
Nama : Suryanto Ang  
NPM : 1205000886  
Program Studi : Ilmu Komputer  
Judul Skripsi : Pengelompokan Dokumen Bahasa Indonesia dengan  
Teknik Reduksi Dimensi Nonnegative Matrix  
Factorization dan Random Projection

**Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Ilmu Komputer pada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia**

### DEWAN PENGUJI

Pembimbing : Dr. Hisar Maruli Manurung, S.Kom ( )  
Penguji : Dra. Mirna Adriani, Ph.D ( )  
Penguji : Dr. Ade Azurat, S.Kom ( )

Ditetapkan di :

Tanggal :

## KATA PENGANTAR

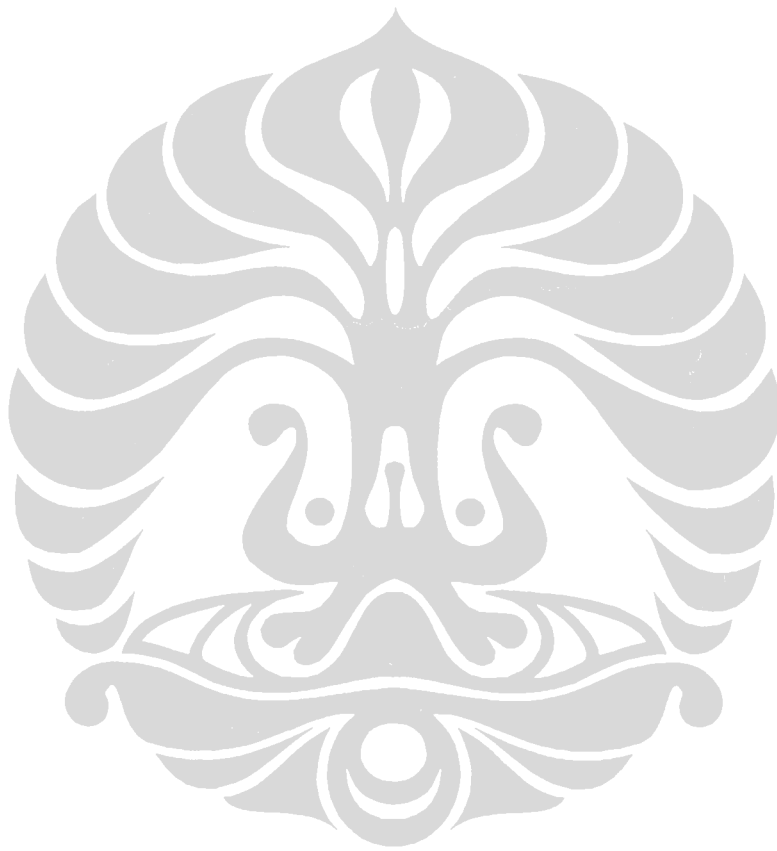
Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan tugas akhir ini dalam jangka waktu yang telah ditetapkan. Atas rahmat dan karunia-Nya juga, penulis dapat menyusun laporan tugas akhir yang berisi penjelasan mengenai tugas akhir yang penulis lakukan. Penulis juga ingin menyampaikan rasa terima kasih kepada semua pihak yang secara langsung maupun tidak langsung telah memberikan bantuan, dukungan serta semangat, sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik. Penulis ingin menyampaikan rasa terima kasih kepada:

1. Kedua orang tua penulis yang selalu mendoakan, memberikan dukungan serta semangat pada setiap kegiatan yang penulis lakukan.
2. Saudara-saudara penulis yang selalu memberikan dukungan dan motivasi sehingga penulis dapat mengerjakan tugas akhir ini dengan semangat.
3. Bapak Hisar Maruli Manurung selaku dosen pembimbing tugas akhir yang telah membimbing penulis dalam pengerjaan tugas akhir hingga penyusunan laporan.
4. Bapak Achmad Nizar Hidayanto selaku pembimbing akademis.
5. Rekan-rekan penulis di Laboratorium Information Retrieval yang telah membantu penulis dan selalu bersama penulis dalam setiap suka dan duka.
6. Rekan-rekan penulis di Fakultas Ilmu Komputer yang telah memberikan bantuan kepada penulis.
7. Semua pihak yang tidak bisa penulis sebutkan disini satu per satu yang telah ikut membantu penulis dalam pengerjaan tugas akhir ini.

Penulis menyadari bahwa penulisan laporan ini tidak sepenuhnya sempurna, masih terdapat kekurangan. Oleh karena itu, penulis sangat mengharapkan saran dan masukan atas laporan ini agar penulis dapat memperbaiki kesalahan pada kesempatan yang lain. Penulis berharap semoga laporan ini dapat memberikan manfaat bagi para pembaca semua.

Jakarta, Juni 2009

**Suryanto Ang**



**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR  
UNTUK KEPENTINGAN AKADEMIS**

---

---

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

Nama : Suryanto Ang  
NPM : 1205000886  
Program Studi : Ilmu Komputer  
Fakultas : Ilmu Komputer  
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 :

PENGELOMPOKAN DOKUMEN BAHASA INDONESIA DENGAN TEKNIK  
REDUKSI DIMENSI NONNEGATIVE MATRIX FACTORIZATION DAN  
RANDOM PROJECTION

Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/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 :

Pada tanggal :

Yang menyatakan

( )

## DAFTAR ISI

HALAMAN JUDUL .....	i
HALAMAN PERNYATAAN ORISINALITAS .....	ii
HALAMAN PENGESAHAN .....	iii
KATA PENGANTAR .....	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS .....	vi
ABSTRAK .....	vii
DAFTAR ISI .....	ix
DAFTAR TABEL .....	xii
DAFTAR GAMBAR .....	xiii
BAB I PENDAHULUAN .....	1
1.1. Latar Belakang .....	1
1.2. Permasalahan .....	2
1.3. Tujuan .....	2
1.4. Ruang Lingkup .....	3
1.5. Metodologi Penelitian .....	3
1.6. Sistematika Penulisan .....	4
BAB II LANDASAN TEORI .....	5
2.1. Pengelompokan Dokumen atau Document Clustering .....	5
2.2. Teknik untuk Mengelompokkan Dokumen .....	7
2.3. Vector Space Model .....	8
2.4. Metrik untuk Pengukuran Tingkat Kesamaan .....	11
2.5. Nonnegative Matrix Factorization .....	13
2.6. Random Projection .....	17
2.7. K-Means .....	20
BAB III PERANCANGAN .....	22
3.1. Alur Pengelompokan Dokumen .....	22
3.2. Data .....	23
3.3. Persiapan Data .....	24
3.4. Penentuan Fitur .....	25
3.5. Term-Document Matrix .....	26



3.6. Teknik Pengelompokan Dokumen .....	28
3.6.1. Nonnegative Matrix Factorization .....	29
3.6.2. Random Projection .....	29
3.7. Pemetaan Kluster yang Dibangun ke Kluster yang Sebenarnya .....	30
3.8. Evaluasi Kinerja .....	31
3.9. Perancangan Eksperimen .....	32
<b>BAB IV IMPLEMENTASI .....</b>	<b>34</b>
4.1. Persiapan Data .....	34
4.2. Implementasi Penentuan Fitur .....	39
4.3. Pembuatan Term-Document Matrix .....	42
4.4. Implementasi Teknik Pengelompokan Dokumen .....	45
4.4.1. Implementasi Teknik Non-Negative Matrix Factorization .....	46
4.4.2. Implementasi Teknik Random Projection dengan K-Means .....	49
4.5. Evaluasi Kinerja .....	51
<b>BAB V HASIL DAN PEMBAHASAN .....</b>	<b>52</b>
5.1. Percobaan Pengelompokan Dokumen .....	52
5.2. Percobaan dari Aspek Fitur .....	58
5.2.1. Analisa Efek Stopwords .....	60
5.2.2. Analisa Pengaruh Jumlah Fitur yang Digunakan .....	62
5.2.3. Analisa Pengaruh Jenis Informasi Fitur .....	63
5.3. Percobaan dari Aspek Parameter Khusus Teknik .....	64
5.3.1. Analisa Parameter Teknik Nonnegative Matrix Factorization .....	67
5.3.2. Analisa Parameter Teknik Random Projection .....	68
5.4. Percobaan dari Aspek Dokumen .....	69
5.4.1. Analisa Pengaruh Jumlah dan Ukuran Kluster .....	71
5.4.2. Analisa Pengaruh Keseragaman Ukuran Kluster .....	74
5.5. Percobaan dari Aspek Teknik Pengelompokan .....	76
5.6. Percobaan dari Aspek Kemiripan Kluster .....	77
5.7. Percobaan dari Aspek Sumber Dokumen .....	81
5.8. Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai .....	82
5.9. Rangkuman Hasil Percobaan .....	85

BAB VI PENUTUP .....	87
6.1. Rangkuman .....	87
6.2. Kesimpulan .....	88
6.3. Kendala .....	88
6.4. Saran .....	89
DAFTAR PUSTAKA .....	90
Lampiran A: Daftar Stopwords .....	92
Lampiran B: Contoh Artikel .....	93
B.1. Contoh Artikel Kompas Kategori Bisnis Keuangan .....	93
B.2. Contoh Artikel Kompas Kategori Olahraga .....	93
B.3. Contoh Artikel Kompas Kategori Kesehatan .....	93
B.4. Contoh Artikel Kompas Kategori Perempuan .....	94
B.5. Contoh Artikel Kompas Kategori Sains .....	94
B.6. Contoh Artikel Kompas Kategori Travel .....	94
B.7. Contoh Artikel Kompas Kategori Properti .....	95
B.8. Contoh Artikel Kompas Kategori Politik Hukum .....	95
B.9. Contoh Artikel Antara Kategori Olahraga .....	96
Lampiran C: Hasil Eksperimen .....	97
C.1. Hasil Eksperimen dari Aspek Fitur: Penggunaan Stopwords, Jumlah Fitur yang Digunakan, Jenis Informasi Fitur .....	97
C.2. Hasil Eksperimen dari Aspek Parameter Khusus Teknik .....	99
C.3. Hasil Eksperimen dari Aspek Jumlah dan Ukuran Kluster .....	100
C.4. Hasil Eksperimen dari Aspek Keseragaman Ukuran Kluster .....	102
C.5. Hasil Eksperimen dari Aspek Teknik: Nonnegative Matrix Factorization, Random Projection dengan K-Means, dan K-Means .....	103
C.6. Hasil Eksperimen dari Aspek Kemiripan Kluster .....	103
C.7. Hasil Eksperimen Pengelompokan Dokumen dari Sumber Berbeda .....	105
C.8. Hasil Eksperimen Pengelompokan Dokumen ke dalam Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai .....	106

## DAFTAR TABEL

Tabel 5.1. Variabel Percobaan .....	55
Tabel 5.2. Akurasi Hasil Percobaan dari Aspek Fitur: Jenis Informasi Fitur, Persentase Fitur yang Digunakan, dan Penggunaan Stopwords dengan Teknik Nonnegative Matrix Factorization .....	59
Tabel 5.3. Akurasi Hasil Percobaan dari Aspek Fitur: Jenis Informasi Fitur, Persentase Fitur yang Digunakan, dan Penggunaan Stopwords dengan Teknik Random Projection .....	60
Tabel 5.4. Efek <i>Stopwords</i> pada Akurasi NMF dan RP .....	60
Tabel 5.5. Pengaruh Jumlah Fitur dan Jenis Informasi Fitur yang Digunakan pada Akurasi NMF dan RP .....	62
Tabel 5.6. Akurasi Hasil Percobaan dari Aspek Nilai Lambda dan Jumlah Iterasi Teknik Nonnegative Matrix Factorization .....	65
Tabel 5.7. Akurasi Hasil Percobaan dari Aspek Jumlah Pengurangan Dimensi dan Tipe Distribusi Matriks Acak teknik Random Projection .....	66
Tabel 5.8. Pengaruh Jumlah dan Ukuran Kluster pada Akurasi NMF .....	72
Tabel 5.9. Pengaruh Jumlah dan Ukuran Kluster pada Akurasi RP .....	72
Tabel 5.10. Pengaruh Keseragaman Ukuran Kluster pada Akurasi NMF dan RP .....	74
Tabel 5.11. Perbandingan Akurasi Pengelompokan dengan NMF, RP dengan K-Means, dan K-Means .....	77
Tabel 5.12. Pengaruh Kemiripan Kluster pada Akurasi NMF .....	79
Tabel 5.13. Pengaruh Kemiripan Kluster pada Akurasi RP .....	79
Tabel 5.14. Akurasi Pengelompokan Dokumen dari Sumber Berbeda .....	82
Tabel 5.15. Hasil Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai dengan Teknik NMF .....	83
Tabel 5.16. Hasil Percobaan Pengelompokan Dokumen ke Kluster yang Jumlahnya Melebihi Jumlah Kategori yang Dipakai dengan Teknik RP .....	84

## DAFTAR GAMBAR

Gambar 2.1. Contoh <i>Term-document Matrix</i> .....	10
Gambar 2.2. Ilustrasi Vektor Dokumen dalam Sistem Koordinat .....	12
Gambar 3.1. Alur Pengelompokan Dokumen .....	22
Gambar 3.2. <i>Term-document Matrix</i> .....	26
Gambar 4.1. <i>Pseudocode</i> Pengambilan Data Artikel dari <i>Website</i> .....	36
Gambar 4.2. Tampilan <i>Website</i> Kompas Kategori Kesehatan .....	36
Gambar 4.3. Tampilan Salah Satu Artikel Kompas .....	37
Gambar 4.4. <i>Pseudocode</i> Penentuan Fitur .....	40
Gambar 4.5. <i>Pseudocode</i> Penghapusan <i>Stopwords</i> .....	41
Gambar 4.6. <i>Pseudocode</i> Pembuatan <i>Term-document Matrix</i> .....	44
Gambar 4.7. <i>Term-document Matrix</i> dengan Informasi Fitur <i>Presence</i> .....	45
Gambar 4.8. Vektor Label Kategori Dokumen .....	45
Gambar 4.9. Tahapan Pengelompokan Dokumen .....	46
Gambar 4.10. <i>Pseudocode</i> Implementasi Nonnegative Matrix Factorization ...	48
Gambar 4.11. <i>Pseudocode</i> Implementasi Random Projection dengan K-Means .....	50
Gambar 4.12. <i>Pseudocode</i> Evaluasi Kinerja .....	51
Gambar 5.1. Grafik Pengaruh <i>Stopwords</i> pada Akurasi NMF dan RP .....	61
Gambar 5.2. Grafik Pengaruh Jumlah Fitur dan Jenis Informasi Fitur pada Akurasi NMF dan RP .....	63
Gambar 5.3. Grafik Percobaan NMF dengan Informasi Jumlah Iterasi dan Konvergensi .....	66
Gambar 5.4. Grafik Pengaruh Nilai <i>Lambda</i> dan Jumlah Iterasi pada Akurasi NMF .....	67
Gambar 5.5. Pengaruh Tipe Distribusi Matriks Acak dan Jumlah Pengurangan Dimensi pada Akurasi RP .....	69

Gambar 5.6. Grafik Pengaruh Jumlah dan Ukuran Kluster pada Akurasi NMF .....	73
Gambar 5.7. Grafik Pengaruh Jumlah dan Ukuran Kluster pada Akurasi RP .	74
Gambar 5.8. Grafik Pengaruh Keseragaman Kluster pada Akurasi NMF dan RP .....	75
Gambar 5.9. Grafik Perbandingan Akurasi Pengelompokan dengan NMF, RP dengan K-Means, dan K-Means .....	77
Gambar 5.10. Grafik Pengaruh Kemiripan Kluster pada Akurasi NMF .....	80
Gambar 5.11. Grafik Pengaruh Kemiripan Kluster pada Akurasi RP .....	80
Gambar 5.12. Akurasi Pengelompokan Dokumen dari Sumber Berbeda .....	82



