DETEKSI KEMIRIPAN TOPIK PROPOSAL JUDUL TUGAS AKHIR DAN SKRIPSI MENGGUNAKAN LATENT SEMANTIC ANALYSIS DI STMIK BUMIGORA MATARAM

I Putu Hariyadi[1], Hartarto Junaedi[2]
(1) STMIK Bumigora Mataram, putu.hariyadi@stmikbumigora.ac.id
(2) Sekolah Tinggi Teknik Surabaya, hartarto.j@gmail.com

ABSTRACT

Research in university has important role in contributing to national development. By knowing the importance of research, students are motivated to be involved in a research which makes contribution to science. Therefore, the appropriateness of research topic taken by students need to be verified. The result of manual verification process is neither efficient, effective, nor accurate. Thus, methods employing Information Technology (IT) are being developed nowadays. This research applied the Latent Semantic Analysis (LSA) method to detect similarity of topic research title. There are 4 steps in applying LSA method; those are preparation, preprocessing, similarity detection and evaluation step. The experimental result using 40 title proposal query for 400 undergraduate final assignments showed that this system is able to detect topic similarity of thesis title proposal with value MAP of 0.8465 on reduced value k=210 with Threshold Cosine similarity of > 0 without DF thresholding. Whereas testing by DF thresholding resulted in MAP value of 0.8744 on reduced value k=270 with threshold cosine similarity > 0.

Key word : latent semantic analysis, enhanced confix stripping stemmer, term weighting, similarity detection, cosine similarity, mean average precision

1. Pendahuluan

Penelitian di perguruan tinggi termasuk STMIK Bumigora memiliki peranan penting untuk menunjang pembangunan nasional, meningkatkan kemajuan bangsa dan daya saing dengan bangsa lainnya. Menyadari peranan tersebut maka mahasiswa didorong untuk melakukan penelitian yang berkontribusi pada ilmu pengetahuan, sehingga kelayakan topik penelitian yang diambil oleh mahasiswa perlu diverifikasi.

Dalam kurun waktu 4 tahun terakhir ini, bagian program studi di STMIK Bumigora Mataram mengalami kesulitan dalam proses memverifikasi kelayakan topik proposal judul Tugas Akhir (TA) dan Skripsi. Verifikasi kelayakan topik ini dilakukan untuk mengetahui kemiripan topik tersebut dengan topik dari TA dan Skripsi mahasiswa yang telah lulus. Berdasarkan verifikasi tersebut akan diperoleh informasi apakah topik tersebut sudah pernah diangkat atau belum, sehingga apabila belum pernah diangkat maka dapat dilanjutkan untuk dikerjakan oleh mahasiswa bersangkutan, namun sebaliknya jika sudah pernah diangkat maka topik tersebut dapat ditolak atau dilakukan pengembangan agar dapat disetujui.

Sampai saat ini proses verifikasi kelayakan topik TA dan Skripsi masih dilakukan secara manual. Bagian program studi melihat proses verifikasi ini tidak efisien dan efektif, serta tidak akurat karena masih ditemukan atau terdapat mahasiswa yang memiliki topik proposal judul yang sama. Hal ini baru diketahui pada pertemuan pertama saat pembimbingan terjadwal mahasiswa tersebut ke dosen pembimbingnya, setelah dinyatakan lolos seleksi awal topik TA dan skripsinya oleh bagian program studi. Selain itu bagian program studi juga melihat permasalahan ini muncul sebagai akibat keterbatasan fasilitas pencarian yang terdapat pada aplikasi *Bumigora Knowledge Center* (BKC) yang digunakan untuk memanajemen sumber daya Ilmu Pengetahuan Teknologi dan Seni (IPTEKS) di STMIK Bumigora Mataram. Fasilitas pencarian yang tersedia pada aplikasi BKC saat ini masih bersifat konvensional yaitu "*exact matching*", sehingga hanya dapat menampilkan sebagian kecil informasi hasil pencarian TA dan skripsi yang dibutuhkan oleh mahasiswa. Hal ini membuat mahasiswa/dosen masih mengalami kesulitan untuk mengetahui apakah ide topik tugas akhir dan skripsi yang rencananya dibuat sudah pernah diangkat oleh mahasiswa lainnya, dan untuk memperoleh informasi TA dan Skripsi yang dapat dijadikan referensi tambahan saat proses pengerjaannya.

Bagian program studi memiliki harapan adanya suatu sistem berbasis Teknologi Informasi (TI) yang dapat mempercepat proses verifikasi atau seleksi kelayakan topik penelitian sehingga dapat menghasilkan topik penelitian yang beragam.

1.1 Tujuan

Adapun tujuan dari penelitian ini adalah menerapkan LSA untuk mendeteksi kemiripan topik proposal TA dan Skripsi.

1.2 Tinjauan Pustaka

1.2.1 Information Retrieval

Information Retrieval (IR) adalah pencarian materi (biasanya dokumen) dari sesuatu yang bersifat tidak terstruktur (biasanya teks) untuk memenuhi kebutuhan informasi dari dalam kumpulan data yang besar (biasanya disimpan dalam komputer)[1]. Sistem IR melakukan pemrosesan awal (*preprocess*) terhadap database dan menerapkan metode untuk menghitung relevansi antara dokumen di dalam database yang telah di preprocess dengan *query* pengguna. Setelah pengguna memasukkan query, sistem IR mengubahnya untuk mengekstrak term-term penting yang konsisten dengan term-term yang diekstrak dari dokumen yang telah di preprocess dan menghitung kemiripan (relevansi) antara query dengan dokumen berdasarkan pada kata-kata tersebut. Sebagai hasilnya, sistem mengembalikan suatu daftar dokumen terurut secara menurun (*descending*) sesuai nilai kemiripannya dengan query pengguna[2].

1.2.2 Inverted Index

Setiap dokumen (termasuk query pengguna) direpresentasikan menggunakan model *bag-of-words* yang mengabaikan urutan dari kata-kata di dalam dokumen. Dokumen diubah ke dalam suatu *bag* berisi kata-kata yang berdiri sendiri. Kata-kata disimpan dalam database pencarian khusus yang dikelola sebagai *inverted index*, yaitu mengubah dokumen asli yang memuat sekumpulan kata-kata ke dalam daftar kata yang berhubungan dengan dokumen terkait dimana kata-kata tersebut muncul. Pembuatan inverted index memerlukan *linguistic processing* yang bertujuan untuk mengekstrak term-term penting dari dokumen yang direpresentasikan sebagai bag of words[2].

Terdapat 5 tahapan dalam pembuatan inverted index yaitu *Markup & Format Removal*, *Tokenization*, *Filtration*, *Stemming*, dan *Weighting*. *Markup & Format Removal* merupakan tahap untuk melakukan penghapusan *tag markup* dan pemformatan khusus dari dokumen seperti HTML. Tokenization merupakan tahap untuk memisahkan kata-kata yang terdapat di dalam kalimat menjadi token atau potongan kata tunggal. Selain itu juga dilakukan penghapusan karakter-karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil *(lower case*) atau disebut dengan *Text Normalizer*. Filtration merupakan proses untuk menentukan term mana yang harus digunakan untuk merepresentasikan dokumen sehingga dapat digunakan untuk menggambarkan isi dari dokumen dan membedakan dokumen tersebut dari dokumen lainnya di dalam koleksi. *Stemming* merupakan proses mereduksi term ke bentuk kata dasar. *Weighting* merupakan tahap untuk melakukan pembobotan pada term[3].

1.2.3 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan pada term. Proses perhitungan bobot melibatkan dua elemen yang saling melengkapi yaitu frekuensi term *i* dalam dokumen *j*, dan invers document frequency dari term *i*. Term yang lebih sering muncul pada dokumen menjadi lebih penting karena dapat mengindikasikan topik dari dokumen. Frekuensi term *i* dalam dokumen *j* didefinisikan sebagai berikut:

$$tf_{ij} = \frac{f_{ij}}{\max_i(f_{ij})} \quad (1)$$

Dimana f_{ij} adalah jumlah kemunculan term *i* pada dokumen *j*. Frekuensi tersebut dinormalisasi dengan frekuensi dari term yang sering muncul pada dokumen.

Inverse document frequency digunakan untuk menunjukkan *discriminative power* dari term *i*. Secara umum term yang muncul di berbagai dokumen kurang mengindikasikan untuk topik tertentu. Rumus dari inverse document frequency didefinisikan sebagai berikut:

$$idf_i = log_2\left(\frac{n}{df_i}\right) \quad (2)$$

Dimana df_i adalah frekuensi dokumen dari term *i* dan dapat diartikan juga sebagai jumlah dokumen yang mengandung term *i*. digunakan untuk meredam efek relatif terhadap tf_{ij}. *Weight* (bobot) W_{ij}dihitung menggunakan pengukuran TF-IDF yang didefinisikan sebagai berikut:

$$W_{ij} = tf_{ij}\ x\ idf_i \quad (3)$$

Bobot tertinggi diberikan kepada term yang sering muncul dalam dokumen *j*, tetapi jarang dalam dokumen lain[2].

1.2.4 DF Thresholding

Document Frequency (DF) thresholding merupakan metode yang paling sederhana untuk mereduksi kata-kata dan mereduksi dimensi vektor. Jumlah dokumen yang memuat fitur dihitung. Hal ini dilakukan pada setiap fitur, sebelum menghapus seluruh fitur dengan frekuensi dokumen lebih kecil dari batas tertentu dan fitur dengan frekuensi lebih besar dari batas tertentu lainnya[4].

1.2.5 Evaluasi IR

Precision dan *recall* merupakan pengukuran yang dihitung menggunakan sekumpulan dokumen yang tidak diranking. Untuk mengevaluasi hasil dari penemuan kembali dokumen yang diranking pada top *k* dokumen yang terambil digunakan *Mean Average Precision (MAP)*. MAP merupakan rata-rata dari nilai average precision untuk sekumpulan query. *Average Precision (AP)* merupakan rata-rata dari nilai precision pada titik atau posisi ranking dimana setiap dokumen relevan yang terambil. Nilai precision untuk dokumen relevan yang tidak terambil adalah 0. MAP didefinisikan sebagai berikut:

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \quad (4)$$

Q adalah kumpulan query $q_j \in Q\{d_1, \ldots d_{mj}\}$. R_{jk} adalah kumpulan dari hasil dokumen terambil yang telah diranking sampai diperoleh dokumen d_k [1].

1.2.6 Latent Semantic Analysis

Latent Semantic Analysis (LSA) adalah sebuah teori dan metode untuk mengekstraksi dan merepresentasikan penggunaan makna kontekstual dari

kata menggunakan komputasi statistik yang diterapkan pada corpus teks yang besar [5]. Secara umum, LSA bekerja pada kumpulan teks yang dari dalamnya dapat didefinisikan kata beserta konteksnya (disebut model semantik)[6].

Analisa struktur *latent semantic* diawali dengan pembentukan *term document matrix* [5]. *Term-document matrix* merupakan matrik yang memuat nilai kemunculan masing-masing kata pada masing-masing dokumen. Pada LSA matrik ini akan digunakan sebagai input awal dari analisa *Singular Value Decomposition (SVD)*. SVD berdasarkan teori aljabar linier yang mengatakan bahwa matrik persegi panjang A dapat dipecah menjadi 3 matrik yaitu matrik orthogonal U, matrik diagonal S, dan transpose dari matrik orthogonal V. Teori ini dapat dinyatakan menggunakan persamaan $A_{mn} = U_{mm} S_{mn} V_{nn}^{T}$ [7].

Teknik SVD pada LSA adalah reduced SVD yang digunakan untuk melakukan proses pengurangan dimensi pada matriks hasil dekomposisi SVD [5]. Pengurangan dimensi bertujuan untuk menemukan makna yang tersembunyi dari kata dan dokumen. Selain itu dengan menggunakan dimensi yang lebih rendah, maka kata-kata yang tidak penting (*noisy*) juga dapat diidentifikasi dan untuk selanjutnya diabaikan (nilainya dikurangi pada matriks)[6].

2. Metodologi

2.1 Analisa Data

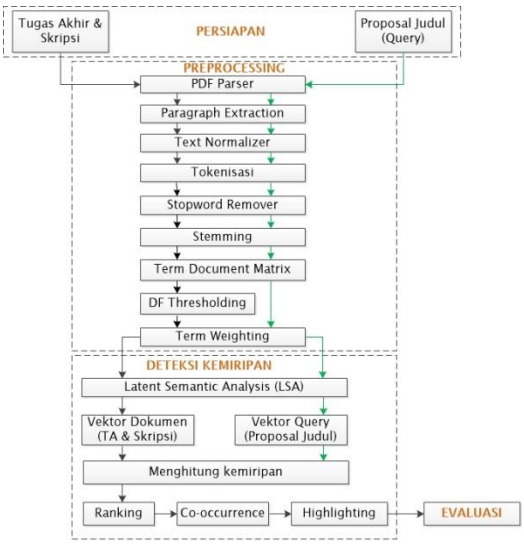
Data penelitian dibagi menjadi dua jenis yaitu data *input* dan data *output*. Data input yang digunakan adalah Tugas Akhir (TA) dan skripsi yang diambil dari aplikasi Bumigora Knowledge Center sejumlah 400, serta proposal judul TA/Skripsi sejumlah 40. Data tugas akhir dan skripsi meliputi judul, abstrak, dan bab 1 khusus bagian latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat. Data proposal judul diinputkan oleh mahasiswa secara mandiri ke dalam sistem. Sedangkan data output yang diharapkan berupa daftar dokumen TA dan Skripsi yang memiliki kemiripan dan persentase kemiripannya dengan proposal TA/Skripsi mahasiswa, serta memperlihatkan detail bagian dari dokumen TA/Skripsi yang memiliki kemiripan yang telah di-highligting dengan proposal tersebut.

2.2 Desain Sistem

Arsitektur dari sistem deteksi kemiripan proposal judul TA dan Skripsi terlihat seperti pada gambar 1. Sistem yang dirancang ini memiliki 4 tahapan yaitu:

A. Tahap Persiapan
   Pada tahap ini dilakukan pengambilan data TA & Skripsi serta proposal judul dan menyimpannya ke database.

B. Tahap *Preprocessing*
   Tahap ini dilakukan baik pada data TA & Skripsi maupun proposal judul mahasiswa, meliputi: *PDF Parser*, *Paragraph Extraction*, Text Normalizer, *Tokenisasi*, *Stopword Remover*, Stemming, *Term document matrix*, DF Thresholding (hanya dilakukan pada skripsi), dan Term Weighting dengan TF-IDF, sebagaimana yang telah dijelaskan pada subbab 2.x. Reduksi fitur dengan DF thresholding dilakukan menggunakan dua batasan yaitu pertama, term yang memiliki TF >=2 dan DF lebih besar dari atau sama dengan setengah dari jumlah dokumen skripsi yang memuat term tersebut. Kedua, term yang memiliki DF sama dengan jumlah dokumen skripsi.



Gambar 1. Arsitektur Sistem

C. Tahap Deteksi Kemiripan
   Tahap ini terdiri dari 4 bagian yaitu pertama, LSA yang digunakan untuk mendeteksi kemiripan proposal judul dengan TA dan Skripsi. Kedua, *cosine similarity* yang digunakan untuk mengukur kemiripan antara vektor query proposal judul dengan vektor dokumen TA dan skripsi sehingga menghasilkan nilai dalam bentuk persentase yang digunakan untuk proses *ranking*. Ketiga, ranking digunakan untuk membuat peringkat hasil dari perhitungan cosine similarity secara menurun (descending). Ke-empat, *highlighting* digunakan untuk menandai bagian dari dokumen TA dan skripsi yang memiliki kemiripan dengan proposal judul menggunakan warna latar belakang tertentu. LSA menggunakan model SVD pada term document matrix skripsi yang telah dilakukan pembobotan menggunakan TF-IDF. Proses dekomposisi pada matrik skripsi akan menghasilkan tiga matrik yaitu matrik *U*, *S*, dan *V*. Setelah mendapatkan SVD maka dilakukan proses reduksi dimensi terhadap matrik U, S dan V sebesar nilai *k*. Untuk menemukan nilai k yang optimal maka sistem dirancang untuk dapat secara dinamis menentukan nilai k yang digunakan sehingga dapat dilakukan percobaan menggunakan beragam nilai k. Hasil reduksi dimensi matrik untuk masing-masing nilai k akan disimpan di database. Selanjutnya dilakukan pencarian koordinat baru dari vektor dokumen dan vector query proposal judul yang direduksi dalam ruang berdimensi k. Baris dari matrik V menyimpan nilai vektor eigen, dimana setiap baris berisi koordinat dari sebuah vektor dokumen. Pengukuran kemiripan dilakukan dengan

cosine similarity antara *vector query* dan vector dokumen, serta hasilnya dirangking.

D. Tahap Evaluasi
Pada tahap ini dilakukan pengukuran kinerja sistem menggunakan MAP. Untuk mendukung proses evaluasi maka dibuat tabel relevansi yang memuat daftar proposal judul dan relevansi terhadap dokumen TA dan skripsi mahasiswa.

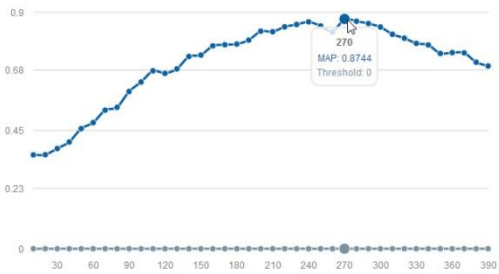
3. Pembahasan

3.1 Kemiripan Dokumen

Pengujian dilakukan menggunakan query 40 proposal judul terhadap 400 skripsi dengan 2 jenis skenario yaitu tanpa reduksi fitur dan dengan reduksi fitur DF thresholding pada tahap preprocessing. Setiap skenario pengujian menggunakan pengaturan parameter-parameter sebagai berikut:

a. Reduksi matrik SVD yang digunakan oleh LSA untuk mendeteksi kemiripan dilakukan dengan berbagai nilai dari parameter k sehingga ditemukan nilai k yang memberikan nilai MAP maksimum. Koleksi dokumen TA/Skripsi adalah 400 maka nilai k maksimum yang dapat dipilih dari term document matrix skripsi adalah 400. Dalam penelitian ini, evaluasi nilai k yang digunakan meliputi 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380 dan 390.
b. Perhitungan cosine similarity dilakukan untuk setiap query terhadap setiap hasil reduksi SVD dengan nilai k=10 sampai dengan k=390.
c. Perhitungan MAP dilakukan dengan berbagai nilai *threshold* terhadap precision dari cosine similarity sehingga ditemukan nilai threshold yang optimal meliputi lebih besar dari 0, 0.05, 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 dan 0.95.

Nilai MAP maksimum diperoleh ketika pengujian dengan reduksi fitur DF thresholding yang menghasilkan nilai MAP sebesar 0.8744 pada nilai reduksi k=270 dengan nilai threshold cosine similarity > 0, seperti terlihat pada gambar 2.



Gambar 2. MAP dengan DF Thresholding

Sedangkan pengujian tanpa reduksi fitur DF thresholding menghasilkan nilai MAP sebesar 0.8465 pada nilai reduksi k=210 dengan nilai threshold cosine similarity > 0, seperti terlihat pada gambar 3.

Gambar 3. MAP tanpa DF Thresholding

3.2 Kemiripan Paragraf

Evaluasi kemiripan paragraf proposal dilakukan menggunakan query berupa paragraf dari proposal terhadap paragraf dari skripsi tertentu dengan dua jenis skenario yaitu tanpa dan dengan DF Thresholding. Ujicoba mengambil salah satu proposal mahasiswa yang memiliki 6 paragraf dengan salah satu skripsi yang memiliki 32 paragraf dengan nilai reduksi k=20, hasil persentase kemiripan tertinggi antar paragrafnya terlihat seperti pada tabel 1.

Tabel 1. Persentase Kemiripan Paragraf Tertinggi

Bagian – Paragraf Proposal	Bagian – Paragraf Skripsi (Persentase Kemiripan)	
	Tanpa DF	Dengan DF
Judul	Bab I - 13 (66.80%)	Bab I-13 (76.20%)
Sinopsis-1	-	-
Sinopsis-2	Bab I - 5 (87.60%)	Bab I - 5 (84.02%)
Sinopsis-3	Bab I - 13 (92.20%)	Bab I - 13 (89.64%)
Sinopsis-4	Bab I - 6 (61.17%)	Bab I - 13 (75.87%)
Sinopsis-5	Bab I - 13 (77.23%)	Bab I - 13 (86.97%)

3.3 Perbandingan Vector Space Model (VSM) dengan LSA

Nilai MAP digunakan sebagai komponen untuk membandingkan VSM dengan LSA, seperti terlihat pada tabel 2.

Tabel 2. MAP VSM dengan LSA

Tanpa DF Thresholding		Dengan DF Thresholding	
VSM	LSA	VSM	LSA
0.8478	0.8465	0.8479	0.8744

Berdasarkan tabel diatas terlihat bahwa penggunaan seleksi fitur DF Thresholding pada LSA dapat meningkatkan nilai MAP, sebaliknya pada VSM tidak secara signifikan meningkatkan nilai MAP.

3.4 Kelebihan dan Kelemahan LSA

Adapun kelebihan dari LSA berdasarkan hasil ujicoba adalah:

a. Mereduksi dimensi matrik.
b. Menggunakan kemunculan kata secara bersamaan (co-occurrence) untuk menemukan kesamaan makna antar kata.

Sedangkan kelemahan LSA adalah:

a. Membutuhkan sumber daya prosesor dan memori yang tinggi untuk pemrosesan terutama pada dimensi matrik berukuran besar.
b. Penentuan nilai k yang optimal untuk reduksi matrik SVD dilakukan secara *trial and error*.

4. Kesimpulan

Adapun kesimpulan yang dapat diambil berdasarkan analisa terhadap hasil pengujian yang telah dilakukan adalah sebagai berikut:

a. Metoda LSA dan Cosine Similarity dapat digunakan untuk mendeteksi kemiripan proposal judul dengan Tugas Akhir dan Skripsi.
b. Nilai MAP tertinggi yaitu 0.8744 diperoleh ketika pengujian menggunakan term document matrix dengan DF thresholding pada reduksi dimensi SVD k=270 dan threshold cosine similarity > 0.
c. Reduksi fitur menggunakan DF thresholding dapat mengurangi ukuran term document matrix skripsi dan meningkatkan nilai MAP.

Adapun saran-saran untuk pengembangan penelitian ini lebih lanjut adalah sebagai berikut:

a. Melakukan eksperimen lebih jauh untuk mengetahui pengaturan parameter-parameter yang dapat meningkatkan nilai MAP.
b. Mengembangkan metode yang dapat mempercepat waktu proses pembentukan term document matrix skripsi.
c. Memperbaiki *library* sastrawi dan menyempurnakan algoritma *Enhanced Confix Stripping Stemmer* karena masih ditemukan kesalahan stemming.

Daftar Pustaka

[1] Christopher D. Manning, Prabhakar Raghavan, dan Hinrich Schutze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009
[2] Cios Krzysztof J., Witold Pedrycz, Roman W., Swiniarski, dan Lukasz A. Kurgan, *Data Mining A Knowledge Discovery Approach*. Springer, 2007.
[3] Dr. E. Garcia, *Document Indexing Tutorial for Information Retrieval Students and Search Engine Marketers*, Mi Islita.com, 2005.
[4] Oystein Lohre Garnes, *Feature Selection for Text Categorisation*, Norwegian University Science and Technology, 2009
[5] Thomas K Landauer, Peter W. Foltz, dan Darrell Laham, "An Introduction to Latent Semantic Analysis", *Discourse Processes*, hal. 259-284, 1998
[6] Ria Hari Gusmita, dan Ruli Manurung, "Penerapan Latent Semantic Analysis (LSA) untuk Menentukan Kesamaan Makna antara Kata dalam bahasa Inggris dan Kata dalam Bahasa Indonesia", Universitas Mercu Buana, hal.8, 2008
[7] Baker, Kirk. *Singular Value Decomposition Tutorial*. The Ohio State University, hal. 15, 2013