

ISSN 2089-1083



EC-Council



Co-host:



STMIK
primakara

PROSIDING Volume 04

SNATIKA 2017

Seminar Nasional Teknologi Informasi, Komunikasi dan Aplikasinya

Malang, 23 November 2017

diorganisasi oleh:

Lembaga Penelitian dan Pengabdian pada Masyarakat

Sekolah Tinggi Informatika dan Komputer Indonesia

SNATIKA 2017

**Seminar Nasional Teknologi Informasi, Komunikasi dan Aplikasinya
Volume 04, Tahun 2017**

PROGRAM COMMITTEE

Prof. Dr. R. Eko Indrajit, MSc, MBA (Perbanas Jakarta)
Tin Tin Hadijanto (Country Manager of EC-Council)
Dr. Eva Handriyantini, S.Kom, M.MT (STIKI Malang)

STEERING COMMITTEE

Laila Isyriyah, S.Kom, M.Kom
Sugeng Widodo, S.Kom, M.Kom
Daniel Rudiaman S., S.T, M.Kom
Subari, S.Kom, M.Kom
Jozua F. Palandi, S.Kom, M.Kom
Koko Wahyu Prasetyo, S.Kom, M.T.I
Nira Radita, S.Pd., M.Pd.

ORGANIZING COMMITTEE

Diah Arifah P., S.Kom, M.T
Meivi Kartikasari, S.Kom, M.T
Chaulina Alfianti O., S.Kom, M.T.
Eko Aprianto, S.Pd., M.Pd.
Saiful Yahya, S.Sn, M.T.
Mahendra Wibawa, S.Sn, M.Pd
Fariza Wahyu A., S.Sn, M.Sn.
Isa Suarti, S.Kom
Elly Sulistyorini, SE.
Roosye Tri H., A.Md.
Endah Wulandari, SE.
Ahmad Rianto, S.Kom
M. Syafiudin Sistiyanto, S.Kom
Muhammad Bima Indra Kusuma

SEKRETARIAT

Lembaga Penelitian dan Pengabdian Kepada Masyarakat
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) – Malang
SNATIKA 2017
Jl. Raya Tidar 100 Malang 65146, Tel. +62-341 560823, Fax. +62-341 562525
Website: snatika.stiki.ac.id
Email: snatika2017@stiki.ac.id

KATA PENGANTAR

Bapak/Ibu/Sdr. Peserta dan Pemakalah SNATIKA 2017 yang saya hormati, pertama-tama saya ucapkan selamat datang atas kehadiran Bapak/Ibu/Sdr, dan tak lupa kami mengucapkan terimakasih atas partisipasi dan peran serta Bapak/Ibu/Sdr dalam kegiatan ini.

SNATIKA 2017 adalah Seminar Nasional Teknologi Informasi, Komunikasi dan Aplikasinya yang diselenggarakan oleh STIKI Malang bekerjasama dengan EC-COUNCIL, APTIKOM Wilayah 7 dan Forum Dosen Kota Malang serta Perguruan Tinggi selaku Co-host: Universitas Nusantara PGRI Kediri dan STMIK Primakara Denpasar-Bali. Sesuai tujuannya SNATIKA 2017 merupakan sarana bagi peneliti, akademisi dan praktisi untuk mempublikasikan hasil-hasil penelitian, ide-ide terbaru mengenai Teknologi Informasi, Komunikasi dan Aplikasinya. Selain itu sesuai dengan tema yaitu "*Keamanan Informasi untuk Ketahanan Informasi Kota Cerdas*", topik-topik yang diambil disesuaikan dengan kompetensi dasar dari APTIKOM Wilayah 7 yang diharapkan dapat mensinergikan penelitian yang dilakukan oleh para peneliti di bidang Informatika dan Komputer. Semoga acara ini bermanfaat bagi kita semua terutama bagi perkembangan ilmu dan teknologi di bidang teknologi informasi, komunikasi dan aplikasinya.

Akhir kata, kami ucapkan selamat mengikuti seminar, dan semoga kita bisa bertemu kembali pada SNATIKA yang akan datang.

Malang, 20 November 2017
Panitia SNATIKA 2017

Daniel Rudiaman S., S.T, M.Kom

**SAMBUTAN KETUA
SEKOLAH TINGGI INFORMATIKA DAN KOMPUTER INDONESIA (STIKI) MALANG**

Yang saya hormati peserta Seminar Nasional SNATIKA 2017,

Puji & Syukur kita panjatkan kepada Tuhan Yang Maha Esa, atas terselenggarakannya Seminar Nasional ini sebagai rangkaian kerjasama dengan EC-COUNCIL, APTIKOM Wilayah 7 dan Forum Dosen Kota Malang serta Perguruan Tinggi selaku Co-host: Universitas Nusantara PGRI Kediri dan STMIK Primakara Denpasar-Bali. Kami ucapkan selamat datang kepada peserta Seminar Nasional serta rekan-rekan perguruan tinggi maupun mahasiswa yang telah berpartisipasi aktif sebagai pemakalah maupun peserta dalam kegiatan seminar nasional ini. Konferensi ini merupakan bagian dari 10 Flag APTIKOM untuk meningkatkan kualitas SDM ICT di Indonesia, dimana anggota APTIKOM khususnya harus haus akan ilmu untuk mampu memajukan ICT di Indonesia.

Konferensi ICT bertujuan untuk menjadi forum komunikasi antara peneliti, penggiat, birokrat pemerintah, pengembang sistem, kalangan industri dan seluruh komunitas ICT Indonesia yang ada didalam APTIKOM maupun diluar APTIKOM. Kegiatan ini diharapkan memberikan masukan kepada *stakeholder* ICT di Indonesia, yang meliputi masyarakat, pemerintah, industri dan lainnya, sehingga mampu sebagai penggerak dalam memajukan ICT Internasional.

Akhir kata, semoga forum seperti ini dapat terus dilaksanakan secara periodik sesuai dengan kegiatan tahunan APTIKOM. Dengan demikian kualitas makalah, maupun hasil penelitian dapat semakin meningkat sehingga mampu bersinergi dengan ilmuwan dan praktisi ICT internasional.

Sebagai Ketua STIKI Malang, kami mengucapkan terimakasih kepada semua pihak atas segala bantuan demi suksesnya acara ini.

“Mari Bersama Memajukan ICT Indonesia”

Malang, 20 November 2017
Ketua STIKI,

Dr. Eva Handriyantini, S.Kom, M.MT.

DAFTAR ISI

| | | Halaman | |
|---|---|--|---------|
| | Halaman Judul | ii | |
| | Kata Pengantar | iii | |
| | Sambutan Ketua STIKI | iv | |
| | Daftar Isi | v | |
| 1 | <i>Erri Wahyu Puspitarini</i> | Analisa <i>Technological Content Knowledge</i> dengan menggunakan <i>Structural Equation Modeling</i> | 1 - 5 |
| 2 | <i>Ina Agustina, Andrianingsih, Ambi Muhammad Dzuhri</i> | Sistem Pendukung Keputusan Analisa Kinerja Tenaga <i>Marketing</i> Berbasis WEB Dengan Menggunakan Metode TOPSIS | 6 - 14 |
| 3 | <i>Ahmad Bagus Setiawan, Juli Sulaksono</i> | Sistem Pendataan Santri Berdasarkan Tingkat Pendidikan di Pondok Pesantren Al-Ishlah Bandar Kidul Kota Kediri | 15 – 18 |
| 4 | <i>Risa Helilintar, Siti Rochana, Risky Aswi Ramadhani</i> | Sistem Pakar Diagnosis Hepatitis Menggunakan Metode K-NN untuk Pelayanan Kesehatan Primer | 19 - 23 |
| 5 | <i>Mety Liesdiani, Enny Listiawati</i> | Sistem Kriptografi pada Citra Digital Menggunakan Metode Substitusi dan Permutasi | 24 - 31 |
| 6 | <i>Devie Rosa Anamisa, Faikul Umam, Aeri Rachmad</i> | Sistem Informasi Pencarian Lokasi Wisata di Kabupaten Jember Berbasis Multimedia | 32 – 36 |
| 7 | <i>Ardi Sanjaya, Danar Putra Pamungkas, Faris Ashofi Sholih</i> | Sistem Informasi Laboratorium Komputer di Universitas Nusantara PGRI Kediri | 37 – 42 |
| 8 | <i>I Wayan Rustana Putra Yasa, I Gusti Lanang Agung Raditya Putra, I Putu Agus Swastika</i> | Sistem Informasi Geografis Pemetaan Penyakit Kronis dan Demam Berdarah di Puskesmas 1 Baturiti Berbasis Website | 43 - 49 |

| | | | |
|----|---|---|-----------|
| 9 | <i>Ratih Kumalasari Niswatin, Ardi Sanjaya</i> | Sistem Informasi Berbasis Web untuk Klasifikasi Kategori Judul Skripsi | 50 - 55 |
| 10 | <i>Rina Firliana, Ervin Kusuma Dewi</i> | Sistem Informasi Administrasi dan Peramalan Stok Barang | 56 - 61 |
| 11 | <i>Patmi Kasih, Intan Nur Farida</i> | Sistem Bantu Pemilihan Dosen Pembimbing Tugas Akhir Berdasarkan Kategori Pilihan dan Keahlian Dosen menggunakan Naïve Bayes | 62 – 68 |
| 12 | <i>Teguh Andriyanto, Rini Indriati</i> | Rancang Bangun Sistem Informasi Sidang Proposal Skripsi di Universitas Nusantara PGRI Kediri | 69 – 73 |
| 13 | <i>Luh Elda Evaryanti, I Gusti Lanang Agung Raditya Putra, I Gede Putu Krisna Juliharta</i> | Rancang Bangun Sistem Informasi Perpustakaan Berbasis Website pada SMK N 1 Gianyar | 74 – 80 |
| 14 | <i>I Kadek Evayanto, I Gusti Lanang Agung Raditya Putra, I Putu Agus Swastika</i> | Rancang Bangun Sistem Informasi Geografis untuk <i>Monitoring</i> Kependudukan di Desa Ubung Kaja Denpasar | 81 - 87 |
| 15 | <i>I Gusti Ayu Made Widyari, I Gusti Lanang Agung Raditya Putra, I Gede Putu Krisna Juliharta</i> | Rancang Bangun Sistem Informasi Data Siswa Praktik Kerja Lapangan (PKL) Berbasis Web Responsive pada SMK TI Udayana | 88 – 94 |
| 16 | <i>Ni Putu Risna Diana Ananda Surya, I Gede Juliana Eka Putra, I Gede Putu Krisna Juliharta</i> | Rancang Bangun Sistem Informasi Akademik Berbasis Website pada Yayasan Perguruan Raj Yamuna | 95 – 102 |
| 17 | <i>Resty Wulanningrum, Ratih Kumalasari Niswatin</i> | Rancang Bangun Aplikasi Identifikasi Tanda Tangan Menggunakan Ekstraksi Ciri PCA | 103 – 107 |

| | | | |
|----|--|---|-----------|
| 18 | <i>Bimo Hario Andityo, Sasongko Pramono Hadi, Lukito Edi Nugroho</i> | Perancangan SOP Pemilihan Pengadaan Proyek TI Menggunakan Metode <i>E-purchasing</i> di Biro TI BPK | 108 - 114 |
| 19 | <i>Kadek Partha Wijaya, I Gede Juliana Eka Putra, I Gede Putu Krisna Juliharta</i> | Perancangan Sistem Informasi Media Pembelajaran Pramuka Berbasis Mobile Apps di Kwarcab Klungkung | 115 – 120 |
| 20 | <i>Ira Diana Sholihati, Irmawati, Dearisa Glory</i> | Aplikasi Data Mining Berbasis Web Menggunakan Algoritma Apriori untuk Data Penjualan di Apotek | 121 – 126 |
| 21 | <i>Sigit Riyadi, Abdul Rokhim</i> | Perancangan Aplikasi Tanggap Bencana Banjir Berbasis SMS Gateway di Desa Kedawung Wetan Pasuruan | 127 – 132 |
| 22 | <i>Fahrudin Salim</i> | Pengaruh <i>Information Technology Service Management (ITSM)</i> terhadap Kinerja Industri Perbankan | 133 - 137 |
| 23 | <i>Fajar Rohman Hariri, Risky Aswi Ramadhani</i> | Penerapan Data Mining menggunakan <i>Association Rules</i> untuk Mendukung Strategi Promosi Universitas Nusantara PGRI Kediri | 138 - 142 |
| 24 | <i>Johan Ericka W.P.</i> | Penentuan Lokasi <i>Road Side Unit</i> untuk Peningkatan Rasio Pengiriman Paket Data | 143 – 147 |
| 25 | <i>Irmawati, Sari Ningsih</i> | Pendeteksi Redundansi Frase pada Pasangan Kalimat | 148 – 153 |
| 26 | <i>Lilis Widayanti, Puji Subekti</i> | Pendekatan <i>Problem Based Learning</i> untuk Meningkatkan Pemahaman Konsep Mahasiswa Prodi Teknik Informatika | 154 – 160 |
| 27 | <i>Sufi Oktifiani, Adhistya Erna Permanasari, Eko Nugroho</i> | Model Konseptual Faktor-Faktor yang Mempengaruhi Literasi Komputer Pegawai Pemerintah | 161 – 166 |
| 28 | <i>Ervin Kusuma Dewi, Patmi Kasih</i> | Meningkatkan Keamanan Jaringan dengan Menggunakan Model Proses Forensik | 167 - 172 |

| | | | |
|----|--|---|-----------|
| 29 | <i>Aminul Wahib, Witarto Adi Winoto</i> | Menghitung Bobot Sebaran Kalimat Berdasarkan Sebaran Kata | 173 – 179 |
| 30 | <i>Evi Triandini, M Rusli, IB Suradarma</i> | Implementasi Model B2C Berdasarkan ISO 9241-151 Studi Kasus Tenun Endek, Klungkung, Bali | 180 – 183 |
| 31 | <i>Ina Agustina, Andrianingsih, Taufik Muhammad</i> | Implementasi Metode SAW (<i>Simple Additive Weighting</i>) pada Perancangan Sistem Pendukung Keputusan Penerimaan Beasiswa Berbasis Web | 184 – 189 |
| 32 | <i>Danar Putra Pamungkas, Fajar Rohman Hariri</i> | Implementasi Metode PCA dan <i>City Block Distance</i> untuk Presensi Mahasiswa Berbasis Wajah | 190 – 194 |
| 33 | <i>Lukman Hakim, Muhammad Imron Rosadi, Resdi Hadi Prayoga</i> | Deteksi Lokasi Citra Iris Menggunakan Threshold Linear dan Garis Horisontal Imajiner | 195 – 199 |
| 34 | <i>Hendry Setiawan, Windra Swastika, Ossie Leona</i> | Desain Aransemen Suara pada Algoritma Genetika | 200 – 203 |
| 35 | <i>Kartika Rahayu Tri Prasetyo Sari, Hisbuloh Ahlis Munawi, Yosep Satrio Wicaksono</i> | Aplikasi <i>Principal Component Analysis</i> (PCA) untuk Mengetahui Faktor yang Mempengaruhi Stres Kerja Perawat | 204 – 208 |
| 36 | <i>Dwi Harini, Patmi Kasih</i> | Aplikasi Bantu Sistem Informasi dan Rute Rumah Sakit di Kota Kediri dengan <i>Local Based Service</i> (LBS) | 209 – 213 |
| 37 | <i>Diah Arifah P., Daniel Rudiaman S.</i> | Analisa Identifikasi <i>Core Point</i> Sidik Jari | 214 – 219 |
| 38 | <i>Mochamad Subianto, Windra Swastika</i> | Sistem Kontrol Kolaborasi Java Programming dan MySQL pada Raspberry Pi | 220 - 225 |
| 39 | <i>Meme Susilowati, Hendro Poerbo Prasetya</i> | Hasil Analisis Proses Bisnis Sistem Informasi Pembiayaan Akademik sesuai Borang Akreditasi | 226 – 230 |

| | | | |
|----|--|---|-----------|
| 40 | <i>Mochamad Bilal, Teguh Andrianto</i> | Uji Kinerja Tunneling 6to4, IPv6IP Manual dan Auto | 231 – 235 |
|----|--|---|-----------|

Pendeteksi Redundansi Frase pada Pasangan Kalimat

Irmawati¹, Sari Ningsih²

Sistem Informasi

Fakultas Teknologi Komunikasi dan Informatika

Universitas Nasional

¹irmawati@civitas.unas.ac.id, ²sariningsih.arif@yahoo.com

ABSTRAK

Pembuatan dokumen spesifikasi kebutuhan perangkat lunak (SKPL) mengharuskan konsumen untuk menyatakan semua kebutuhan dengan spesifik dan teliti sebelum tahapan perancangan mulai dilakukan. Hal ini dilakukan untuk mencegah atau mengurangi kemungkinan pengulangan dalam tahap perancangan, pengkodean, dan uji coba. Kualitas dokumen spesifikasi kebutuhan yang rendah dapat menyebabkan over budget, keterlambatan dan sistem tidak dapat diandalkan. Salah satu penyebab rendahnya kualitas dokumen spesifikasi kebutuhan adalah pelanggan seringkali mengungkapkan kebutuhannya dalam bahasa yang tidak dimengerti pengembang, sementara pengembang hanya berfokus pada aspek teknis dari pengembangan perangkat lunak. Hal tersebut menimbulkan redundansi informasi kebutuhan sistem sehingga dokumen spesifikasi kebutuhan menjadi tidak konsisten. Penelitian ini mengusulkan suatu solusi untuk mendeteksi redundansi dalam dokumen spesifikasi kebutuhan perangkat lunak berdasarkan kalimat yang dibentuk, serta frase yang sudah terdefiniskan pada kalimat–kalimat sebelumnya. Kami mengusulkan rancangan dan aplikasi pendeteksi redundansi frase pada pasangan kalimat dalam dokumen spesifikasi kebutuhan perangkat lunak berbahasa Inggris dan Indonesia. Teknik yang digunakan untuk menghitung kemiripan kalimat pada frase pada penelitian ini menggunakan metode Vector Space Mode. Hasil dari penelitian ini dapat memberikan kontribusi untuk menghasilkan dokumen spesifikasi kebutuhan perangkat lunak yang berkualitas baik dan sebagai media pembelajaran proses text mining.

Kata Kunci: Redundansi, Frase, Pasangan Kalimat, Vector Space Model

1. Pendahuluan

Penilaian kualitas dokumen spesifikasi kebutuhan perangkat lunak membutuhkan waktu cukup lama jika dilakukan secara manual, dengan panjang kebutuhan dokumen berkisar ratusan halaman yang berisi sejumlah kata, frase dan kalimat yang masing-masing berpotensi untuk disalah artikan. Kualitas dokumen spesifikasi kebutuhan yang rendah dapat menyebabkan over budget, keterlambatan dan sistem tidak dapat diandalkan. Salah satu penyebab rendahnya kualitas dokumen spesifikasi kebutuhan adalah pelanggan seringkali mengungkapkan kebutuhannya dalam bahasa yang tidak dimengerti pengembang, sementara pengembang hanya berfokus pada aspek teknis dari pengembangan perangkat lunak tanpa menyadari bahwa bahasa yang digunakan tidak dimengerti oleh pelanggan. Hal tersebut akan menimbulkan redundansi informasi kebutuhan sistem sehingga dokumen spesifikasi kebutuhan menjadi tidak konsisten. Redundansi pada spesifikasi kebutuhan perangkat lunak terjadi jika

kebutuhan yang sama dinyatakan lebih dari satu kali. Masalah yang disebabkan redundansi terjadi ketika spesifikasi kebutuhan perangkat lunak direvisi, jika semua redundansi tidak dirubah maka spesifikasi kebutuhan perangkat lunak menjadi tidak konsisten. Sedangkan konsistensi merupakan salah satu kategori yang merepresentasikan kualitas yang harus dimiliki dari suatu dokumen spesifikasi kebutuhan perangkat lunak sesuai dengan standar yang sudah ditentukan, salah satu standarnya IEEE Standard 830. Berdasarkan uraian diatas, maka kami mengusulkan suatu solusi untuk mendeteksi redundansi frase dalam dokumen spesifikasi kebutuhan perangkat lunak berdasarkan kalimat yang dibentuk, serta munculnya frase yang sudah terdefiniskan pada kalimat–kalimat sebelumnya. Penelitian ini dilakukan untuk menciptakan suatu aplikasi pendeteksi redundansi frase pada pasangan kalimat dalam dokumen spesifikasi kebutuhan perangkat lunak. Dalam menghitung kemiripan kalimat dan frase pada penelitian ini penulis menggunakan metode *Vector*

Space Model dan *Phrase Affinity Clustering (PAC)*. Keluaran dari penelitian ini adalah desain kerangka kerja pendeteksi redundansi frase pada pasangan kalimat, sebuah aplikasi pendeteksi redundansi frase pada pasangan kalimat dalam dokumen spesifikasi kebutuhan perangkat lunak yang berbahasa Inggris dan bahasa Indonesia dan aplikasi pendeteksi redundansi berbasis web yang dapat digunakan sebagai media pembelajaran *text mining* karena didukung dengan kemampuan menunjukkan proses *text mining* berupa Proses *tokenizing, filtering* dan *stemming*. Kontribusi pada penelitian ini adalah meningkatnya dokumen spesifikasi kebutuhan perangkat lunak yang berkualitas baik, menurunkan tingkat kegagalan suatu proyek pengembangan perangkat lunak dan dapat menjadi media pembelajaran *text mining*.

2. Metode Penelitian

Penelitian yang telah dilakukan sebelumnya oleh para peneliti mengenai deteksi redundansi pada pasangan kalimat dan dokumen digunakan untuk menunjang penelitian ini. [1] mengajukan kerangka kerja pendeteksi redundansi pasangan kalimat pada dokumen spesifikasi kebutuhan perangkat lunak. Hasil pengujian dan evaluasi didapatkan nilai Kappa yaitu 0,747 dan nilai threshold 0,58. Sehingga solusi yang diajukan dapat digunakan untuk mendeteksi redundansi pada pasangan kalimat, karena berdasarkan interpretasi nilai Kappa, kekuatan proporsi kesepakatan antara solusi yang diajukan dengan ahli bahasa adalah banyak (*substantial agreement*). [2] dalam penelitiannya mengajukan metode minipar untuk parsing ketergantungan dan pohon penyesuaian bottom up untuk penyesuaian pasangan snippet, untuk mendeteksi dan menghapus informasi-informasi redundansi menggunakan algoritma *graph-based representation* di tingkat sub-snippet pada dokumen berita. Hasil penelitian yang dilakukan oleh [3] membuat sistem penggabungan kalimat otomatis untuk menggabungkan beberapa kalimat masukan menjadi sebuah kalimat baru dengan metode *intersection fusion*. Terdapat empat tahapan dalam penggabungan kalimat, yaitu: konstruksi pohon, penjajaran antar pohon, penggabungan pohon, dan pembangkitan kalimat gabungan. Sistem dikembangkan dengan menggunakan Stanford Parser untuk konstruksi pohon dan Java Wordnet Interface untuk mencari sinonim antara dua frase.

Penelitian yang dilakukan [4] menghasilkan *phrase structure tree* dan memodifikasi hasilnya dengan komponen berbasis aturan sehingga menjadi pohon dependensi. Pohon dependensi yang digunakan pada penelitian ini menyimpan fitur dari kata yang ada dari tiap simpul. Tujuan dari penyimpanan fitur simpul tersebut untuk memudahkan pembangkitan kata dalam tahap selanjutnya dan kumpulan kata dapat membentuk frase kata benda dan kata kerja dengan partikel dijadikan satu buah simpul frase. Pendeteksian kemiripan antar frase menggunakan WordNet dan sebuah kamus *paraphrase*, pada tahap ini pasangan frase pada kalimat berbeda dinilai kemiripannya dengan menggunakan WordNet. Jika relasi antar frase tersebut tidak ditemukan maka digunakan kamus *paraphrase*. Penelitian yang menghasilkan algoritma terbaru untuk clustering [5], pada penelitian tersebut pengelompokan dilakukan pada dokumen text yang dikembangkan berdasarkan *vector space model*, frase dan algoritma *affinity propagation clustering (APC)*. Algoritma baru yang dihasilkan dinamakan *Phrase Affinity Clustering (PAC)*. Proses pertama pada algoritma *Phrase Affinity Clustering (PAC)* yaitu menemukan frase dengan algoritma *ukkonen suffix tree construction*, proses kedua menemukan model ruang vector menggunakan skema pembobotan *tf-idf* pada frase dan proses ketiga menghitung matrik kemiripan dengan *cosine similarity*.

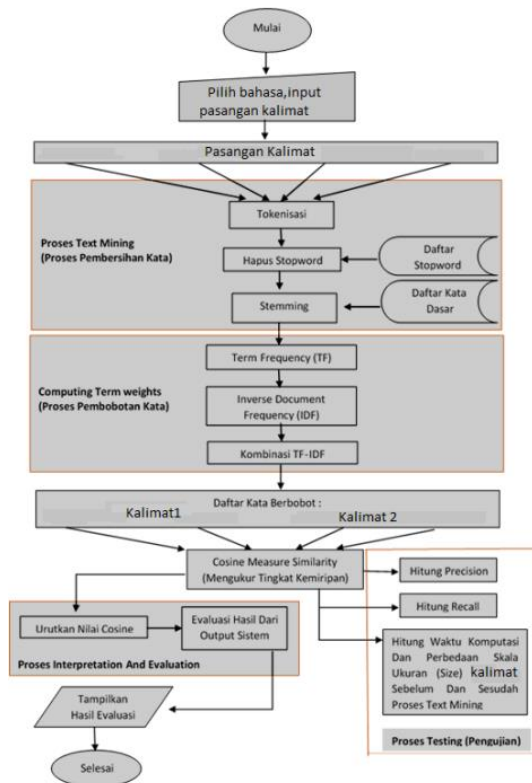
3. Tahapan Penelitian

Pada tahap rancangan produk dibuat satu buah modul dengan tiga buah submodul untuk mendukung penelitian. Modul dan submodul tersebut adalah:

Modul pengindeksan dengan tiga submodul:

1. Submodul proses *Text Mining*.
2. Submodul proses perhitungan pembobotan frase (*Tf-Idf*).
3. Submodul pengukuran tingkat kemiripan (*measure similarity*).

Flowchart keseluruhan sistem pendeteksi redundansi frase dapat dilihat pada Gambar berikut:



Gambar 1. Flowchart keseluruhan sistem

Tahap-tahap dimulai dari pemrosesan data yang tidak terstruktur menjadi terstruktur sampai pada penyaringan data hingga ditemukan sebuah *knowledge* atau relevansi hasil informasi yang diperlukan *user* oleh sistem, dijabarkan sebagai berikut:

1. *User* diharuskan untuk memasukan pasangan kalimat, penggunaan bahasa (Indonesia dan Inggris) yang akan dibandingkan.
2. Sistem menyimpan pasangan kalimat yang telah dimasukan oleh user kedalam basis data.
3. Proses *Text Mining*, pada tahap ini dilakukan teknik *Text Mining* pada pasangan kalimat yang telah didapat dari langkah sebelumnya yang kemudian akan dibersihkan dan dipersiapkan untuk tahap selanjutnya. Proses untuk mempersiapkan kalimat meliputi proses pembersihan kalimat dari karakter-karakter yang tidak diperlukan, proses penghapusan *stopword* (kata penghubung) dan proses *stemming*.

Proses *Text Mining* meliputi proses Tokenisasi, penghapusan *Stopword* dan proses *Stemming*:

1. Tokenisasi

Proses penghilangan tanda baca pada kalimat sehingga menghasilkan kata-kata yang berdiri sendiri [6].

2. Penghapusan *Stopword*

Pengambilan kata-kata yang penting dari hasil tokenisasi. Tahap ini dapat menggunakan algoritma *stoplist* atau *wordlist*. *Stoplist* yaitu penyaringan (*filtering*) terhadap kata-kata umum yang tidak mengacu kepada suatu hal, seperti kata depan, kata gabung, dan sebagainya.

3. Proses *Stemming*

Proses merubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda [6]. *Stem* (akar kata) adalah bagian dari kata yang tersisa setelah dihilangkan imbuhan (awalan dan akhiran).

4. Metode *Vector Space Model* (VSM)

Metode untuk melihat tingkat kedekatan atau kesamaan (*similarity*) *term* dengan cara pembobotan *term*. Dokumen dipandang sebagai sebuah *vector* yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model*, sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah kalimat dengan kalimat lainnya didasarkan pada similaritas diantara vektor kalimat yang satu dengan kalimat lainnya. Dalam VSM koleksi kalimat direpresentasikan sebagai sebuah *matrix term document* (matrik *term frequency*). Setiap sel dalam matrik bersesuaian dengan bobot yang diberikan dari suatu *term* dalam dokumen yang ditentukan. Nilai nol berarti bahwa *term* tersebut tidak ada dalam kalimat. Metode pembobotan dilakukan menggunakan algoritma *Term Frequency – Inverse Document Frequency* (TF-IDF).

Computing term weights (perhitungan bobot dari *term*)

1. *Term Frequency* (*tf*)

Mencari frekuensi kemunculan kata dari pasangan kalimat pada setiap kalimat dengan perhitungan *term frequency* (*tf*) menggunakan persamaan (1).

$$tf = tf_{ij} \dots(1)$$

Dimana, tf adalah *term frequency* dan tf_{ij} adalah banyaknya kemunculan term t_i dalam dokumen d_j . *Term Frequency* (tf) dihitung dengan menghitung banyaknya kemunculan *term* t_i dalam dokumen d_j .

2. *Document Frequency dan Inverse Document Frequency (idf)*

Merupakan logaritma dari jumlah kata di setiap kalimat keseluruhan dibagi dengan jumlah kalimat yang ada. Perhitungan *Inverse Document Frequency* (idf) menggunakan persamaan (2).

$$idf_i = \log(N/df_i) \quad \dots(2)$$

Dimana, idf_i adalah *inverse document frequency*, N adalah jumlah semua kalimat dalam koleksi, df_i adalah banyaknya jumlah kalimat yang mengandung term t_i . Perhitungan idf_i digunakan untuk mengetahui banyaknya term yang dicari (df_i) yang muncul dalam kalimat yang ada pada korpus.

3. Pembobotan Menggunakan $tf-idf$

Perhitungan bobot kata dilakukan dengan mengalikan *Term Frequency* pada setiap dokumen dengan *Inverse Document Frequency* (TF-IDF). Perhitungan TF-IDF menggunakan persamaan (3).

$$W_{ij} = tf_{ij} \cdot \log(N/df_i) \quad \dots(3)$$

Dimana, W_{ij} adalah bobot term t_i terhadap kalimat d_j , tf_{ij} adalah jumlah kemunculan term t_i dalam kalimat d_j , N adalah jumlah semua kalimat yang ada dalam database (korpus), df_i adalah jumlah kalimat yang mengandung term t_i (minimal ada satu kata yaitu term t_i). Bobot kalimat W_{ij} dihitung untuk didapatkannya suatu bobot hasil perkalian atau kombinasi antara *term frequency* tf_{ij} dan *Inverse Document Frequency* idf_i .

Setelah didapatkan nilai hasil dari pembobotan pada tiap kata di setiap kalimat, dalam proses ini dilakukan perhitungan pengukuran nilai tingkat kemiripan dengan membandingkan antara kedua vektor yang bersesuaian. Untuk mengimplementasikan metode

ruang vektor, diasumsikan sudah tersedia sekumpulan *term* yang dapat mendeskripsikan kumpulan kalimat yang tersimpan dalam suatu sistem. Nilai tingkat kemiripan *query-document* bisa didapatkan dengan membandingkan antara kedua vektor yang bersesuaian menggunakan metode Ruang Vektor dengan persamaan *cosine similarity*. Sebuah *vector matching* dapat melakukan perbandingan global antara vektor *query* dan kalimat untuk mendapatkan tingkat *similarity* antara keduanya [7].

Berdasarkan hasil perhitungan bobot dengan TF-IDF yang telah dilakukan dan untuk dapat membandingkan antara kedua vektor, maka perlu dilakukan perhitungan jarak jarak kalimat menggunakan persamaan (4).

$$|dj| = \sqrt{(\sum_{i=1}^t (W_{ij})^2)} \quad \dots(4)$$

Dimana, $|dj|$ adalah jarak kalimat, dan W_{ij} adalah bobot kalimat ke- i , maka Jarak kalimat ($|dj|$) dihitung untuk didapatkan jarak kalimat dari bobot pasangan kalimat (W_{ij}) yang terambil oleh sistem. Jarak kalimat bisa dihitung dengan persamaan akar jumlah kuadrat dari kalimat. Dan melakukan perhitungan pengukuran *similaritas* pasangan kalimat menggunakan persamaan (5).

$$Sim(q, dj) = \sum_{i=1}^t W_{iq} \cdot W_{ij} \quad \dots(5)$$

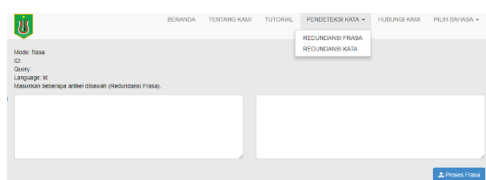
Dimana, W_{ij} adalah bobot *term* dalam kalimat, W_{iq} adalah bobot kalimat, dan $Sim(q, dj)$ adalah *Similaritas* antara kata dan kalimat. *Similaritas* antara kata dan kalimat atau $Sim(q, dj)$ digunakan untuk mendapatkan bobot dengan didasarkan pada bobot *term* dalam kalimat (W_{ij}) dan bobot *query* (W_{iq}) atau dengan cara menjumlah bobot kata (q) dikalikan dengan bobot kalimat. Sehingga pengukuran nilai tingkat kemiripan bisa didapatkan dengan membandingkan antara kedua vektor yang bersesuaian dengan rumus *Cosine Similarity* menggunakan persamaan (6).

$$Sim(q, d_j) = \frac{q \cdot d_j}{|q| \cdot |d_j|} = \frac{\sum_{i=1}^t W_{iq} \cdot W_{ij}}{\sqrt{\sum_{i=1}^t (W_{iq})^2} \cdot \sqrt{\sum_{i=1}^t (W_{ij})^2}} \quad \dots(6)$$

4. Proses *Interpretation and Evaluation*, dalam proses ini pola-pola yang telah diidentifikasi oleh sistem kemudian di terjemahkan/ diinterpretasikan ke dalam bentuk knowledge yang lebih mudah dimengerti oleh *user* untuk membantu dalam mengetahui hasil yang telah diberikan sistem atau bentuk lain yang lebih mudah dimengerti. Penghitungan similaritas pada langkah sebelumnya akan menghasilkan bobot pada tiap kalimat yang menentukan seberapa relevan kalimat tersebut terhadap query, sehingga dapat ditampilkan pasangan kalimat yang relevan saja, secara terurut mulai dari yang paling relevan (bobot tertinggi).
5. Proses *Testing* (pengujian), dalam proses ini dilakukan pengujian terhadap hasil dari setiap proses yang dilakukan sistem. Untuk memperoleh perangkat lunak dengan hasil yang baik, diperlukan pengukuran kualitas terhadap hasil yang didapatkan dari sistem.

4. Hasil dan Pembahasan

Berikut rancangan desain aplikasi pendeteksi redundansi frase pada pasangan kalimat:



Gambar 2. Tampilan halaman pendeteksi frase

Pada tampilan halaman pendeteksi redundansi frase, user diharuskan memilih bahasa Indonesia atau bahasa Inggris yang ingin digunakan dalam pendeteksian, memasukan pasangan kalimat dari dokumen spesifikasi kebutuhan perangkat lunak. Selanjutnya user meng-klik tombol proses untuk mendapatkan hasil dari sistem. Dilakukan pengujian terhadap hasil dari setiap proses yang dilakukan sistem. Untuk memperoleh perangkat lunak dengan hasil yang baik, diperlukan pengukuran kualitas terhadap hasil yang didapatkan dari sistem, yaitu:

1. Pengujian terhadap hasil proses Text Mining dari sistem.

2. Pengujian terhadap hasil proses perhitungan pembobotan kata (Tf-Idf) dan hasil pengukuran tingkat kemiripan (measure similarity) dari sistem.
3. Pengujian terhadap waktu komputasi yang didapat dengan menggunakan stopwatch serta perbedaan skala ukuran (size) dokumen sebelum dan sesudah dilakukan proses Text Mining oleh sistem.
4. Pengujian terhadap hasil evaluasi yang dihasilkan sistem dengan menghitung nilai Recall dan nilai Precision untuk mengetahui tingkat optimal dari hasil sistem. Evaluasi pada sebuah sistem dengan menggunakan Recall dan Precision sudah cukup baik untuk baik untuk menjadi ukuran dari sistem tersebut [7].

uji Recall dan Precision untuk mendapatkan informasi hasil pencarian yang didapatkan oleh sistem yang dibuat. Precision dapat dianggap sebagai ukuran ketepatan atau ketelitian, sedangkan Recall adalah kesempurnaan. Nilai Precision adalah proporsi kalimat yang terambil oleh sistem adalah relevan. Nilai Recall adalah proporsi dokumen relevan yang terambil oleh sistem [9].

Nilai Recall dan Precision bernilai antara 0 sampai dengan 100%. Sistem ini diharapkan dapat memberikan nilai Recall dan Precision mendekati 100% sebagai ketepatan dan kesempurnaan sistem dalam menghasilkan informasi yang relevan [7].

5. Kesimpulan

Tercapainya aplikasi yang dibuat dapat mendeteksi frase yang redundan pada pasangan kalimat dalam bahasa Indonesia dan bahasa Inggris dan terbentuknya kerangka kerja pendeteksi frase pada pasangan kalimat dalam dokumen spesifikasi kebutuhan perangkat lunak.

6. Referensi

- [1] Oktaviani, E.R., Siahaan, D. 2013. ICEIC 2013. *Redundancy Detection In Software Requirements Statements*. Bali, Indonesia, Jan 30 – Feb 2, 2013.
- [2] Tala, F. Z. 2004. *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*, Master Thesis, Universiteit van Amsterdam, The Netherlands.

- [3] Thadani K., McKeown K. 2008. *Framework For Identifying Textual Redundansi*. Proceedings of The 22nd International Conference On Computational Linguistics (Coling 2008). Manchester, 2008.
- [4] Barzilay, Regina; McKeown, Kathleen R. 2005. Sentence Fusion for Multidocument News Summarization. *Journal Computational Linguistics*. vol. 43, pp. 297-328.
- [5] Setiawan, W., Khodra, M. L. 2012. Pengembangan Sistem Penggabung Kalimat Otomatis. *Jurnal Sarjana Institut Teknologi Bandung Bidang Teknik Elektro dan Informatika*. Volume 1 hal. 139-144.
- [6] Arifin, A. Z. dan Setiono, A. N. 2002. *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*. Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA). Surabaya: ITS, 2002
- [7] Wahono, S. R. 2003. Analyzing Requirements Engineering Problems. IECI Japan Workshop 2003 (IJW-2003). Japan, 2003.
- [8] Yates, R.B. 1999. *Modern Information Retrieval, Addison Wesley-Pearson international edition*. Boston. USA.
- [9] Salton Gerard and Christopher Buckley. 1998. Term-Weighting Approaches In Automatic Text Retrieval. *Information Processing & Management*. Vol.24, pp. 513-523.
- [10] Shrivastava, S. K., Rana, J. L., Jain, R. C. 2013. Text Document Clustering based on Phrase Similarity using Affinity Propagation. *International Journal of computer Applications* (0975-8887). volume 61- No.18, January 2013.