Penerapan Algoritma Peningkatan *Porter Stemmer* dan *Likelihood* Untuk Mengidentifikasi Topik Artikel Berita

OLEH :

DEVI ANDRIYANI 1212100088

DOSEN PEMBIMBING :

1. Dr. Imam Mukhlash, S.Si, MT
2. Alvida MustikaRukmi, S.Si, M.Si

OUTLINE

- PENDAHULUAN
- TINJAUAN PUSTAKA
- METODE PENELITIAN
- HASIL DAN PEMBAHASAN
- KESIMPULAN DAN SARAN
- DAFTAR PUSTAKA

PENDAHULUAN

- LATAR BELAKANG

Seiring dengan perkembangan teknologi yang pesat semakin meningkat pula penyebaran informasi secara online seperti halnya berita atau artikel yang mudah sekali kita jumpai pada berbagai situs. Sekumpulan informasi tersebut tentunya memiliki tema pembicaran yang beragam sehingga tidak mungkin semua informasi yang disajikan bisa dicerna secara bersamaan, melainkan harus dikelompokkan berdasarkan relevansi topik dari berita tersebut. Pengelompokan tersebut dapat mempermudah pembaca untuk memilih informasi yang paling penting sesuai dengan topik yang ingin dibaca.

PENDAHULUAN

- RUMUSAN MASALAH

1. Bagaimana merepresentasikan suatu berita / teks dokumen agar metode klasifikasi dokumen dapat diterapkan?
2. Bagaimana menerapkan algoritma peningkatan Porter Stemmer pada proses stemming dokumen?
3. Bagaimana menerapkan metode likelihood untuk membuat aplikasi yang dapat mengidentifikasi topik dokumen?
4. Bagaimana *performance* nilai akurasi hasil identifikasi topik (perbandingan dengan penelitian terdahulu)?

PENDAHULUAN

- BATASAN MASALAH

1. Dokumen yang digunakan adalah berita Bahasa Indonesia
2. Dokumen berita untuk training dan testing menggunakan corpus yang diunduh dari www.kompas.com dan disimpan dalam bentuk notepad yang kemudian di-convert ke eksistensi .news
3. Jumlah dokumen berita dan jumlah kategori primitif yang digunakan sesuai dengan jumlah data yang ada di database.
4. Aplikasi dibuat menggunakan Bahasa pemrograman Java dan database MySQL

PENDAHULUAN

- TUJUAN

1. Membuat aplikasi yang dapat mengidentifikasi topik dari berita berbahasa Indonesia.

2. Mengetahui hasil penerapan algoritma peningkatan Porter Stemmer saat proses stemming dalam hal akurasi hasil identifikasi topik dokumen berita berbahasa Indonesia.

PENDAHULUAN

- MANFAAT

1. Program ini diharapkan nantinya dapat menjadi media penunjang yang dapat mempermudah pengguna dalam memilih informasi dari dokumen berita sesuai topik yang diinginkan, dengan kata lain sebagai alat yang secara otomatis dapat mengidentifikasi topik berita tanpa harus melalui proses manual seperti pada umumnya.
2. Mendapatkan efisiensi waktu dalam pencarian topik pada berita yang ingin dibaca.
3. Memperluas wawasan soal kinerja metode yang paling baik untuk diterapkan.

TINJAUAN PUSTAKA

- *VEKTOR SPACE MODEL*

[4] Vector space model merupakan salah satu pendekatan yang paling umum digunakan untuk merepresentasikan model teks digital. Setiap dokumen dj akan direpresentasikan menjadi vector.

$$dj = (w1j , w2j , . . . , wij) \qquad (1)$$

dimana wij adalah bobot term ke- i pada dokumen j yang bersangkutan.

TINJAUAN PUSTAKA

- METODE TF-IDF

[5] Bobot setiap *term* dapat direpresentasikan dengan frekuensi invers dokumennya (TF-IDF) yang dinyatakan sebagai berikut :

$$w_{ij} = tf_{ij} \log_2 \left(\frac{N}{df_i} \right) \qquad (2)$$

dimana w_{ij} adalah bobot *term* ke- i pada dokumen j yang bersangkutan, tf_{ij} adalah frekuensi term i pada dokumen j. N adalah jumlah dokumen yang diproses dan df_j adalah dokumen yang memiliki *term* i di dalamnya.

TINJAUAN PUSTAKA

- *LIKELIHOOD*

[1] Perhitungan likelihood untuk sebuah kategori dijelaskan pada rumus (3).

$$\text{Likelihood } (c_j \mid A=\{k_1,k_2,\ldots,kn\}) = -\sum_{i=1}^{n} P(k_i|c_j) \log\left(P(k_i|c_j)\right) \qquad (3)$$

dimana c_j adalah kategori, A adalah artikel dokumen uji, $P(k_i|c_j)$ dihitung menggunakan "In-Document" dan perhitungan "jumlah total dokumen".

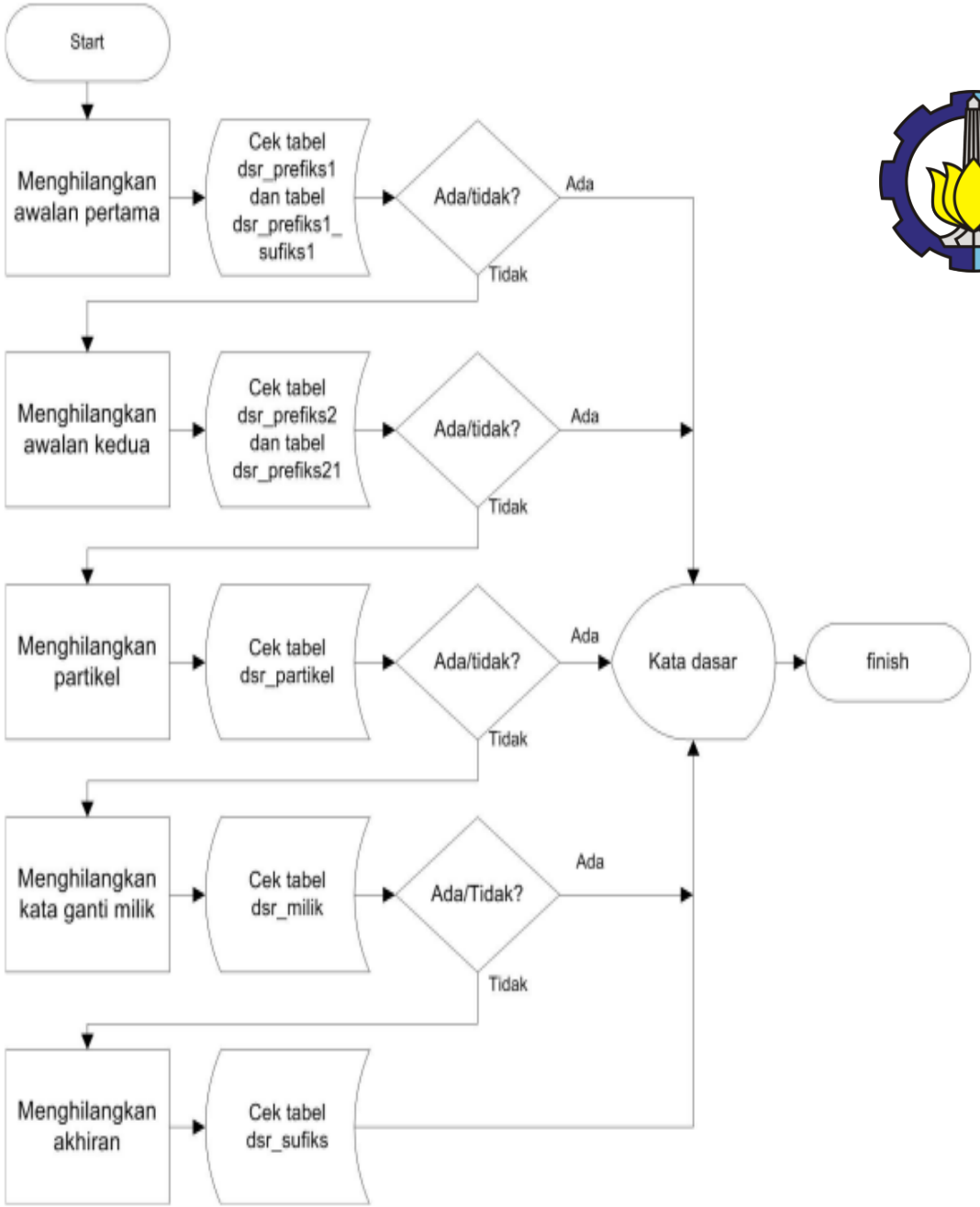
Nilai ambang (threshold) digunakan untuk menentukan apakah sebuah kategori dapat ditetapkan untuk artikel uji atau tidak.

$$Threshold = \frac{\sum_1^{|L|} l_i}{|L|} + \sqrt{\frac{\sum\left(l_i - \frac{\sum_1^{|L|} l_i}{|L|}\right)^2}{|L|}} \qquad (4)$$

dengan L adalah jumlah banyaknya likelihood, l_i adalah likehood untuk kategori ke-i.

TINJAUAN PUSTAKA

- ALGORITMA PENINGKATAN *PORTER STEMMER*



TINJAUAN PUSTAKA

- ALGORITMA IDENTIFIKASI TOPIK

$$CosSim\ (t_i, A) = \frac{t_i A}{|t_i||A|} \qquad (5)$$

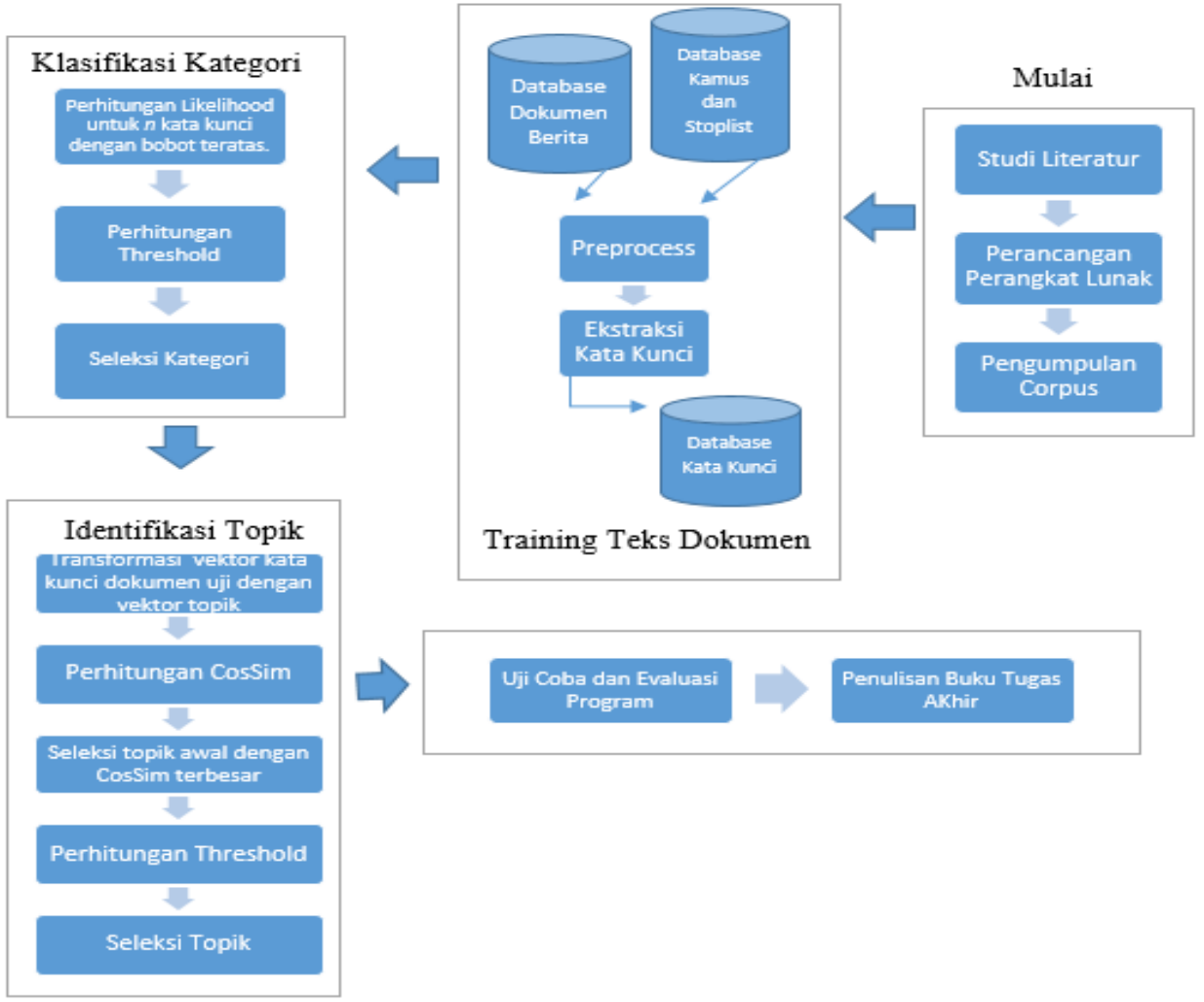
dengan t_i adalah vector topik ke-i, A adalah artikel uji A, $|t_i|$ dan $|A|$ masing-masing adalah panjang vector topik ke –i dan panjang vector artikel A.

$$\text{NewTSim}\ (t_c, A) = \frac{(0.05\ x\ |t_c| x\ (mean(A) - StdDev(A))\ x\ mean(t_c)}{\left(|A| x\ (mean(A))^2\right)\ x\ (|t_c|\ x\ (mean(t_c))^2)} \qquad (6)$$

(i) $CosSim\ (t_c, A) > 0.1\ \wedge CosSim\ (t_c, A) > NewTSim\ (t_c, A)$

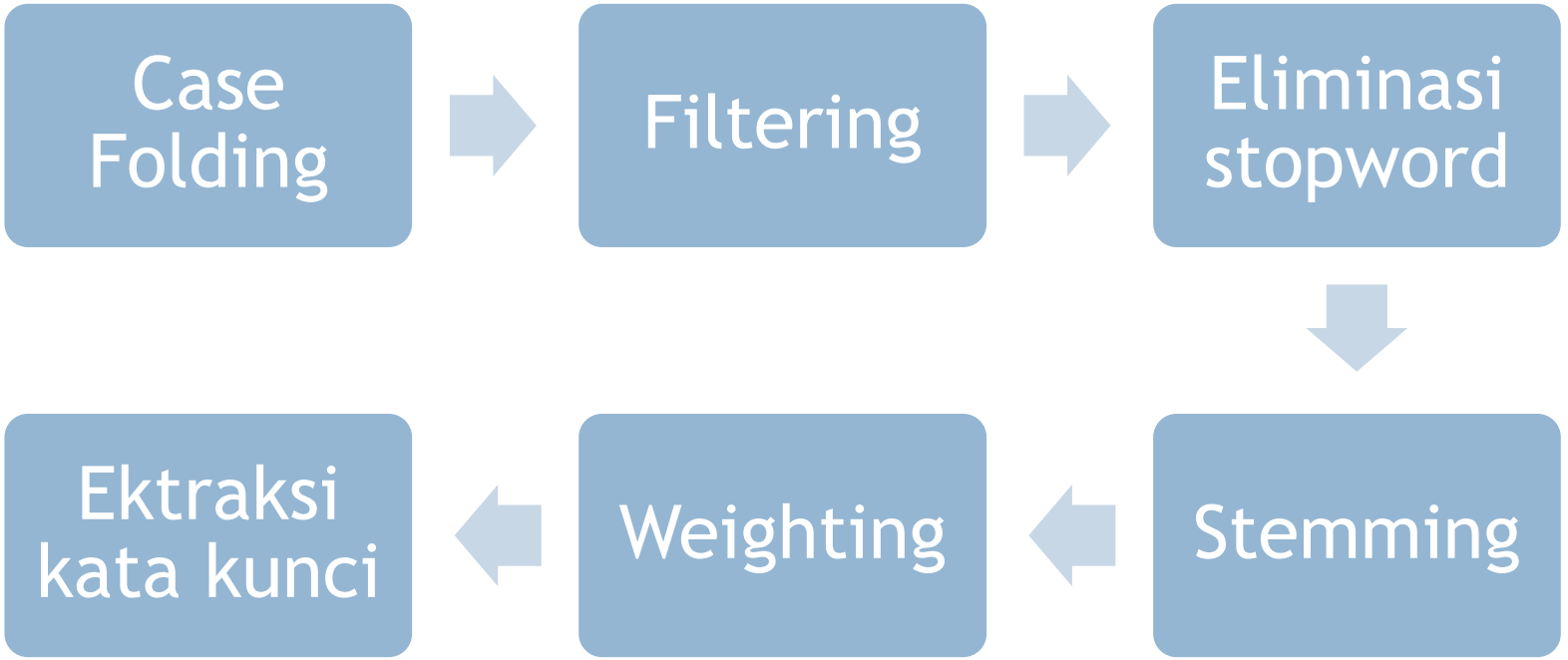
(ii) $NumTopics > 10\ \wedge CosSim\ (t_c, A) > (2\ x\ StdDev(AllTopicSims) + Mean(AllTopicSims))$

METODE PENELITIAN



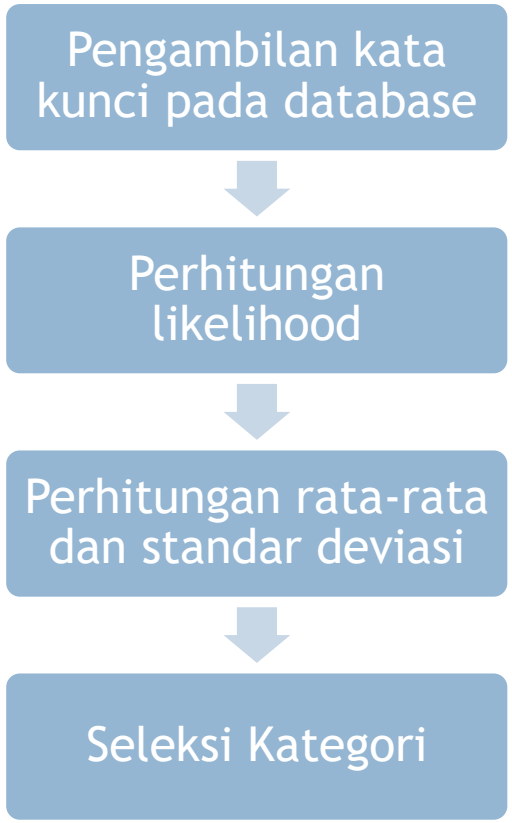
METODE PENELITIAN

- TRAINING



METODE PENELITIAN

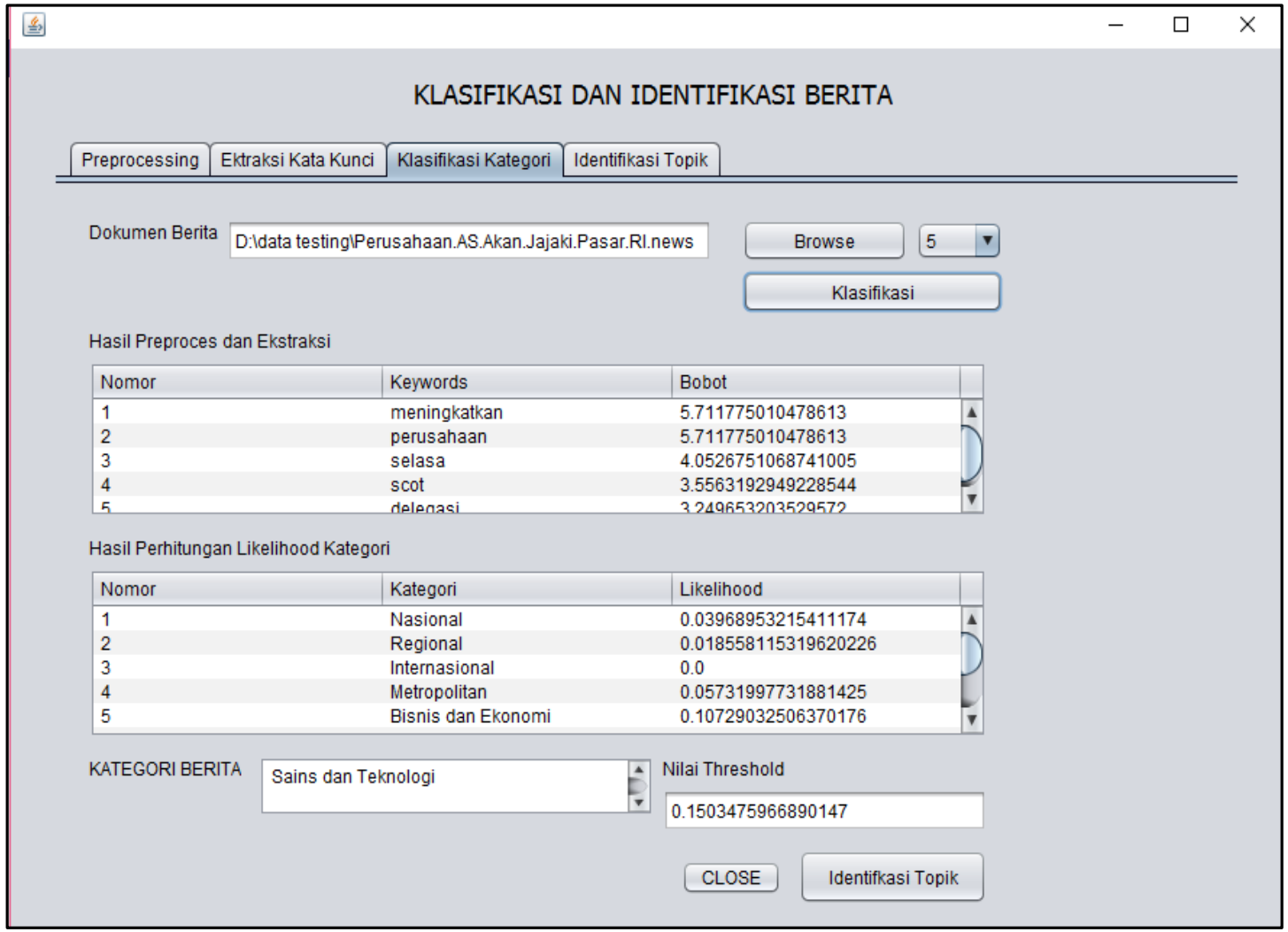
- KLASIFIKASI KATEGORI

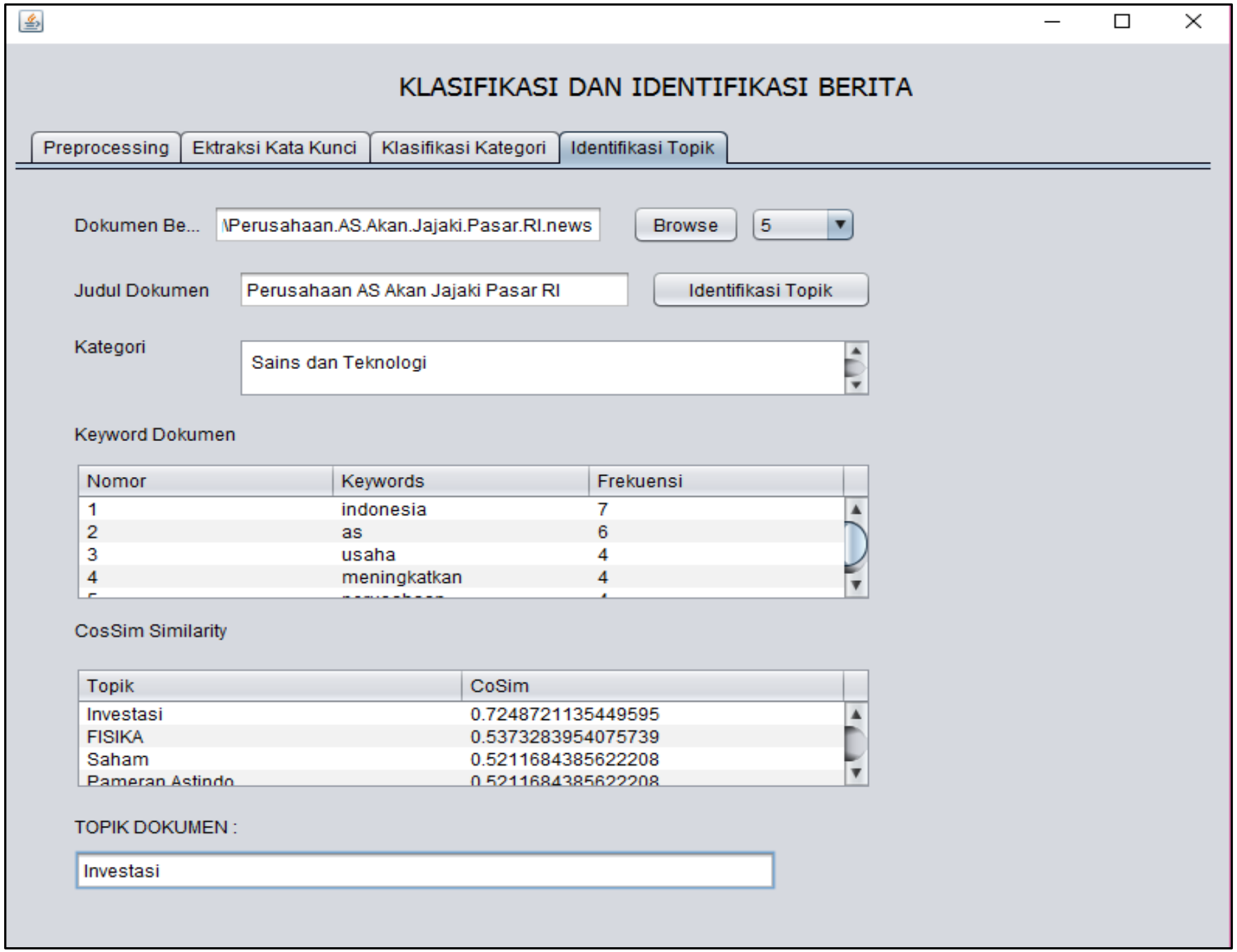


METODE PENELITIAN

- IDENTIFIKASI TOPIK

HASIL DAN PEMBAHASAN




KLASIFIKASI DAN IDENTIFIKASI BERITA
Preprocessing
Ektraksi Kata Kunci
Klasifikasi Kategori
Identifikasi Topik
Dokumen Be...
\Perusahaan.AS.Akan.Jajaki.Pasar.RI.news
Browse
5
Judul Dokumen
Perusahaan AS Akan Jajaki Pasar RI
Identifikasi Topik
Kategori
Sains dan Teknologi
Keyword Dokumen
Nomor
Keywords
Frekuensi
1 indonesia 7
2 as 6
3 usaha 4
4 meningkatkan 4
CosSim Similarity
Topik
CoSim
Investasi 0.7248721135449595
FISIKA 0.5373283954075739
Saham 0.5211684385622208
Pameran Astindo 0.5211684385622208
TOPIK DOKUMEN :
Investasi

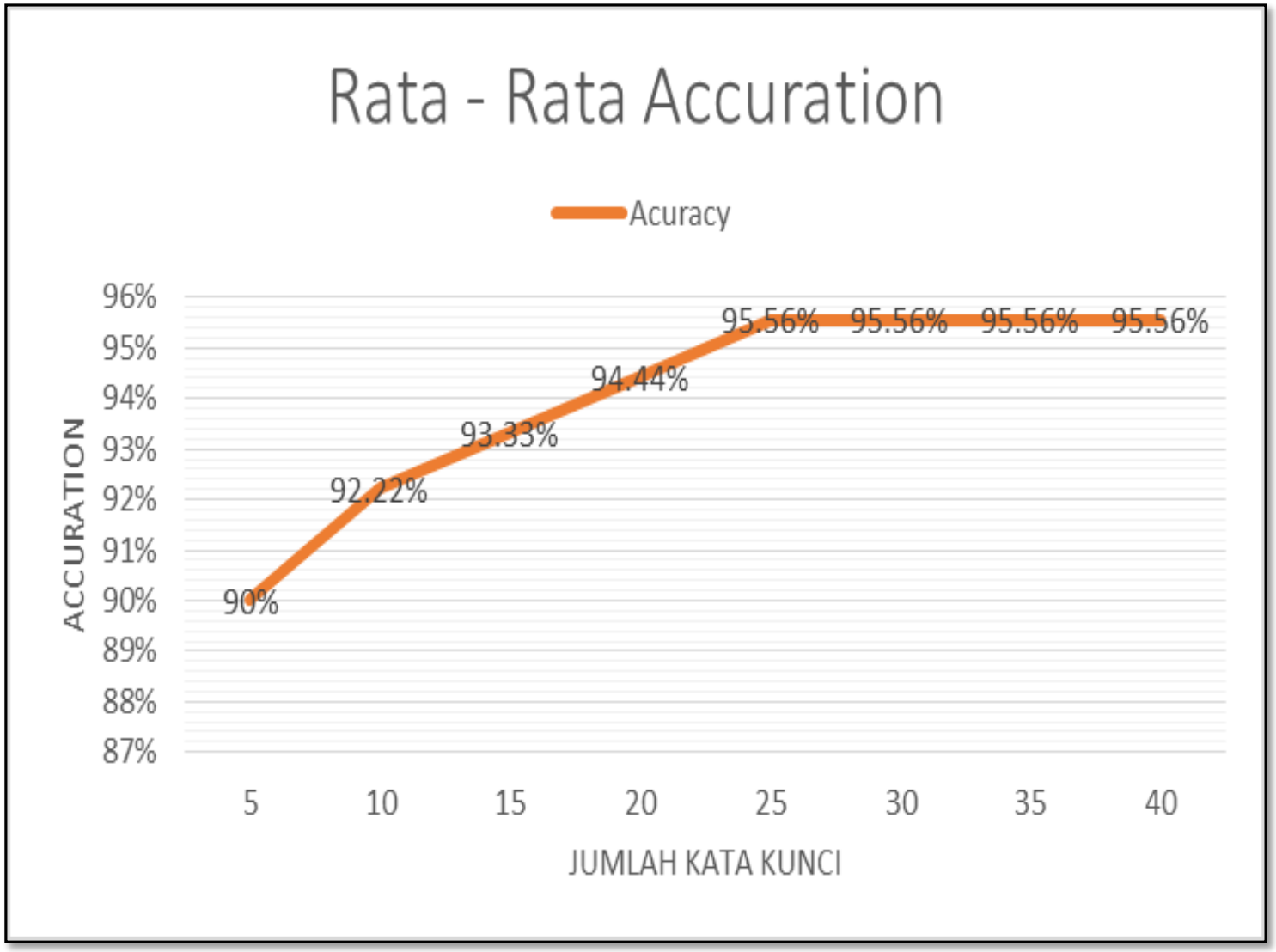
UJI COBA DAN EVALUASI

Kategori	Jumlah Dokumen
Nasional	10
Regional	10
Internasional	10
Metropolitan	10
Bisnis dan Ekonomi	10
Olahraga	10
Sains dan Teknologi	10
Edukasi	10
Pariwisata	10
Total	**90**

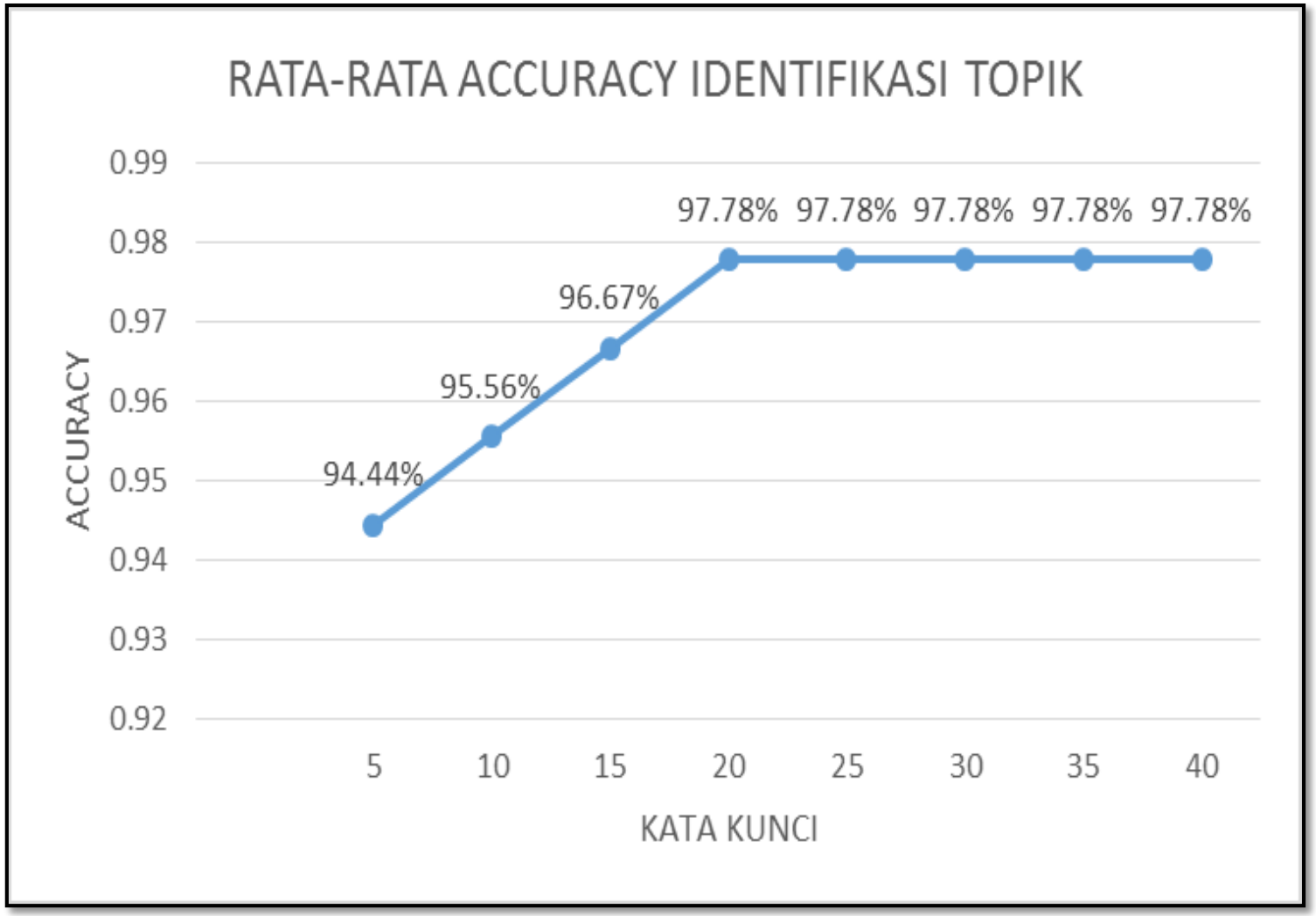
- *Precision* (P) = TP / (TP + FP)
- *Recall* (R) = TP / (TP + FN)
- *F-Measure* (F) = 2 * P * R / (P + R)
- *Accuration* (A) = (TP + TN) / (TP + FP + FN + TN)

ACCURATION KLASIFIKASI KATEGORI

	5	10	15	20	25
	Accuration	Accuration	Accuration	Accuration	Accuration
Kategori					
Internasional	100.00%	100.00%	90.00%	80.00%	100.00%
Nasional	90.00%	100.00%	100.00%	100.00%	100.00%
Regional	80.00%	70.00%	90.00%	90.00%	90.00%
Metropolitan	80.00%	80.00%	80.00%	100.00%	100.00%
Bisnis Ekonomi	90.00%	100.00%	100.00%	100.00%	100.00%
Olahraga	90.00%	90.00%	90.00%	90.00%	90.00%
Pariwisata	90.00%	100.00%	100.00%	100.00%	100.00%
Sains & Teknologi	100.00%	100.00%	100.00%	100.00%	100.00%
Edukasi	90.00%	90.00%	90.00%	90.00%	80.00%
Rata - rata	90.00%	92.22%	93,33%	94,44%	95,56%


Rata - Rata Accuration
Acuracy
96%
95%
94%
93%
92%
91%
90%
89%
88%
87%
ACCURATION
90%
92.22%
93.33%
94.44%
95.56%
95.56%
95.56%
95.56%
5
10
15
20
25
30
35
40
JUMLAH KATA KUNCI

ACCURATION IDENTIFIKASI TOPIK



KESIMPULAN DAN SARAN

Kesimpulan

- Program telah selesai dibuat menggunakan Algoritma Peningkatan Porter Stemmer dan Likelihood serta diuji mampu melakukan proses klasifikasi kategori serta identifkasi topik pada artikel berita berbahasa Indonesia
- Berdasarkan hasil uji coba, proses klasifikasi kategori mendapatkan hasil yang optimal saat menggunakan jumlah kata kunci sebanyak 25, sedangkan untuk identifikasi topik diperoleh hasil yang maksimal dengan jumlah kata kunci sebanyak 20.
- Nilai *accuracy* untuk klasifikasi kategori diperoleh sebesar 95,56 %, sedangkan untuk identifikasi topik sebesar 97,78 %. Kedua nilai tersebut tampak lebih baik daripada nilai *accuracy* yang dihasilkan pada penelitian sebelumnya.

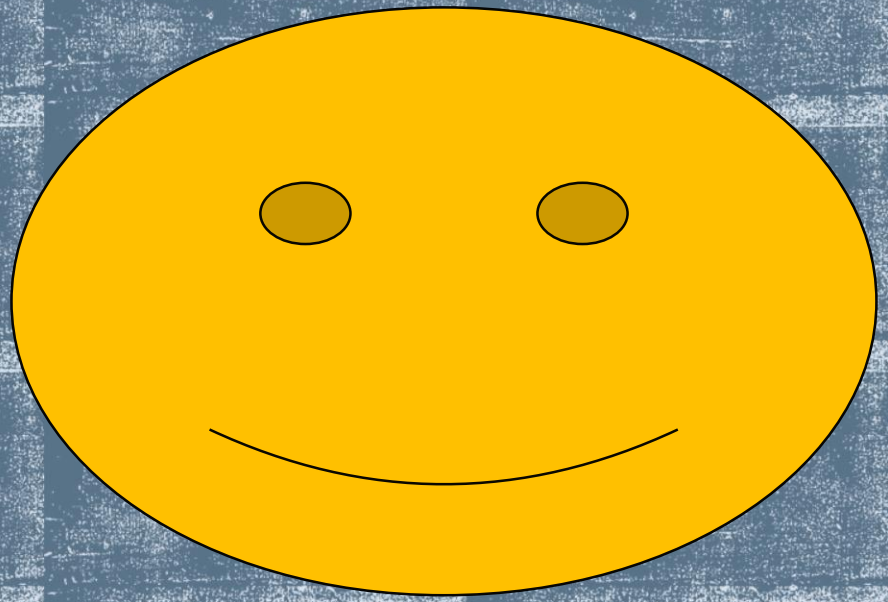
KESIMPULAN DAN SARAN

Saran

- Riset lebih lanjut dalam hal *running time*, karena membutuhkan waktu yang cukup lama saat identifikasi topik.
- Program disediakan fungsi *download* dokumen agar secara otomatis disimpan mengikuti format Corpus.

DAFTAR PUSTAKA

- Bracewell D, Jiajun Yan, Fuji Ren dan Shingo Kuroiwa.2009. “Category Classification and Topic Discovery of Japanese and English News Article,” Electronic Notes in Theoretical Computer Science 225(2009) 51-65
- Fuddoly, Aini Rachmania Kusumaagama, Agus Zainal Arifin.2011. “Klasifikasi Kategori dan Identifikasi Topik pada Artikel Berita Bahasa Indonesia,” ITS.Surabaya
- Karaa,Wahiba Ben Abdessalem, “A New Stemmer to Improve Information Retrieval,” *International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013*
- DR.E. Garcia,2006. The Classic Vector Space Model, URL:http://www.miislita.com/term-vector/term-vector-3.html
- Wiguna ,Putu Bagus Susastra, Bimo Sunarfri Hantono."Peningkatan Algoritma Porter Stemmer Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran," JNTETI, Vol.2, No.2,2013
- Agusta Ledy, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia", Konferensi Nasional Sistem dan Informatika 2009; Bali, November 14, 2009
- Nadirman Firnas, 2006. **Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model Pada Pencarian File Dokumen Berbasis Teks**, <URL:http://kabulkurniawan.web.ugm.ac.id/wp-content/uploads/SKRIPSI.pdf>



TERIMA KASIH