

**IDENTIFIKASI KECOCOKAN DOKUMEN TANYA JAWAB
MENGUNAKAN METODE *TEXT MINING* DAN *VECTOR
SPACE MODEL*.**

SKRIPSI



Oleh:

**Masti Fatchiyah Maharani
1434010077**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR
2018**

LEMBAR PENGESAHAN
SKRIPSI

Judul : IDENTIFIKASI KECOCOKAN DOKUMEN TANYA JAWAB
MENGUNAKAN METODE *TEXT MINING* DAN *VECTOR
SPACE MODEL*

Oleh : MASTI FATCHIYAH MAHARANI
NPM : 1434010077

Telah Diseminarkan Dalam Ujian SKripsi Pada:
Hari Senin, Tanggal 21 Mei 2018


Dosen Pembimbing


Menyetujui

Dosen Penguji

1.


1.



Fetty Tri Anggraeny, S.Kom, M.Kom
NPT : 3 82020 6020 81


Budi Nugroho, S.Kom, M.Kom
NPT : 3 8009 050 205 1


2.

2.


Intan Yuniar P., S.Kom, M.Sc
NPT : 3 8006 04019 81


Eva Yulia P., S.Kom, M.Kom
NPT : 3 8907 130 346 1

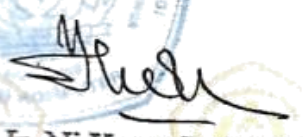
3.

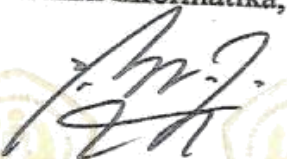

Mohammad Idhom S.P., S.Kom, MT
NPT : 3 8303 100 285 1

Mengetahui :

Dekan
Fakultas Ilmu Komputer,

Koordinator Program Studi
Teknik Informatika,


Dr. Ir. Ni Ketut Sari, MT
NIP. 19650731 199203 2001


Budi Nugroho, S.Kom, M.Kom
NPT. 380090502051

IDENTIFIKASI KECOCOKAN DOKUMEN TANYA JAWAB MENGUNAKAN METODE TEXT MINING DAN ALGORITMA VECTOR SPACE MODEL (VSM).

Nama Mahasiswa : Masti Fatchiyah Maharani
NPM : 1434010077
Program Studi : Teknik Informatika
Dosen Pembimbing : Fetty Tri Anggaraeny S.Kom, M.Kom
Intan Yuniar P. Skom, M.sc

Abstrak

Information Retrieval (IR) atau sistem temu kembali merupakan sebuah sistem yang bekerja untuk menemukan informasi yang relevan dengan apa yang dicari oleh pengguna. Dalam penelitian yang dilakukan oleh penulis, proses *IR* diterapkan dengan menggunakan metode *text mining* dan algoritma *vector space model*. *Text mining* merupakan proses awal dari alur kerja sebuah sistem temu kembali, yang mana proses ini bertujuan untuk menghasilkan kumpulan kata-kata yang berhubungan dengan dokumen yang dicari. Proses *IR* kemudian dilanjutkan dengan proses mencari nilai kemiripan antar dokumen yang relevan menggunakan algoritma *vector space model*. Algoritma *vector space model* merupakan sebuah algoritma yang menggambarkan sebuah dokumen sebagai sebuah bentuk *vector* yang memiliki jarak dan arah.

Dalam penerapannya, algoritma ini melalui beberapa tahapan seperti *indexing* dokumen yang dilakukan dalam proses *text mining*, pembobotan term yang dalam penelitian ini dilakukan dengan menggunakan algoritma TF-IDF, dan perhitungan kesamaan. Ada beberapa jenis metode yang digunakan dalam menghitung kesamaan dokumen dalam algoritma *vector space model*, salah satunya adalah *cosine similarity* yang digunakan dalam penelitian ini. *Cosine similarity* mengukur kesamaan antar dua vektor dengan mengambil nilai kosinus dari sudut antara vektor kata kunci dan vektor tiap dokumen.

Dari hasil penelitian ini, diketahui bahwa sistem mampu menghasilkan nilai rata-rata *precision* sebesar 3.04%

Kata kunci: *Information Retrieval (IR)*, *Text Mining*, *Vector Space Model(VSM)*.

KATA PENGANTAR

Segala puji bagi Allah SWT yang Maha pengasih lagi Maha penyayang sebab karena karunia, kebaikan, dan kemudahannya lah penulis dapat menyelesaikan skripsi yang berjudul “Identifikasi Kecocokan Dokumen Tanya Jawab Menggunakan Metode *Text Mining* dan *Vector Space Model*” sesuai dengan rentang waktu yang direncanakan.

Skripsi ini dibuat sebagai salah satu persyaratan untuk memenuhi salah satu syarat dalam menempuh ujian sarjana pada Fakultas Ilmu Komputer, Program Studi Teknik Informatika di Universitas Pembangunan Nasional “Veteran” Jawa Timur.

Penulis berharap bahwa dengan penyusunan skripsi ini mampu membuka “gerbang” ilmu baru dan memberikan manfaat bagi teman-teman pembaca. Namun, penulis juga menyadari bahwa dalam penyusunan skripsi ini masih terdapat banyak kesalahan maka dari itu penulis berharap adanya saran dan kritik membangun bagi penulis dan juga semoga penelitian ini dapat dikembangkan dan disempurnakan kedepannya.

Surabaya, Mei 2018

Penulis

UCAPAN TERIMA KASIH

Dalam penyusunan dan penulisan skripsi ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu dalam kesempatan ini penulis dengan senang hati menyampaikan terima kasih kepada yang terhormat:

1. Bapak Prof. Dr. Ir. H. Teguh Soedarto, M.P selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Ibu Dr. Ir. Ni Ketut Sari, M.T. selaku Dekan Fakultas Ilmu Komputer.
3. Bapak Budi Nugroho, S.kom., M.kom., selaku Ketua Progdi Teknik Informatika.
4. Ibu Fetty Tri Anggraeny, S.kom., M.kom., selaku pembimbing I dan Ibu Intan Yuniar Purbasari, S.Kom, M.Sc., selaku pembimbing II, yang dengan sabar, dan ikhlas meluangkan waktu, tenaga dan pikiran memberikan bimbingan, motivasi, arahan, dan saran- yang sangat berharga kepada penulis selama menyusun skripsi.
5. Staff Dosen Teknik Informatika UPN “Veteran” Jawa Timur yang telah membekali penulis dengan berbagai ilmu selama mengikuti perkuliahan sampai akhir penulisan skripsi
6. (Alm) Ayah, Bunda, Abang, dan Mbak yang selalu memberikan dukungan secara moril dan materiil dalam penyusunan skripsi ini
7. Partner skripsi penulis yang telah membantu banyak dalam penyusunan konsep dan pengerjaan skripsi, Khoritul Arif Yanto.

8. Teman-teman yang bersedia mendengarkan keluh kesah serta menghibur penulis dikala jenuh yaitu Widya Sekar Arum dan David Risdianto
9. Teman-teman Teknik Informatika angkatan 2014, khususnya kelas B, yang sudah bersedia mengisi dan menjadi bagian dalam cerita perkuliahan penulis.
10. Teman-teman asisten laboratorium algoritma dan pemrograman yang memberikan banyak pengalaman berharga dalam sisi lain perkuliahan yakni Vivi, Andi, Irul, dan Raharjo.
11. Teman-teman BEM Fakultas Ilmu Komputer 2017-2018
12. Teman-teman UK Pers Mahasiswa, khususnya pengurus 2016, Hani, Arum, Kholis, Ilham, Radit, Inggrid, Silvi, Bachtiar, Fitri, dan Miya yang sudah banyak memberikan cerita semasa kuliah.
13. Teman-teman KKN 2017, khususnya group SSI, Stefany, Sonia, Rica, Sam, dan Fau yang juga turut memberi dorongan kepada penulis untuk menyelesaikan skripsi ini.
14. Tika, Linda, & Yasina, teman-teman baik saya yang selalu hadir dalam segala kondisi termasuk menyemangati penulis agar dapat menyelesaikan skripsi ini.
15. Bebe dan Mbak Dai, tim kamseupay, yang selalu memberi energi-energi positif dan mendukung penulis dalam segala keadaan.
16. Semua pihak yang tidak dapat penulis ucapkan yang juga turut mendoakan dan menyemangati penulis dalam menyelesaikan skripsi ini.

DAFTAR ISI

LEMBAR PENGESAHAN	Error! Bookmark not defined.
Abstrak	i
KATA PENGANTAR	ii
UCAPAN TERIMA KASIH.....	iii
DAFTAR ISI.....	v
DAFTAR GAMBAR, GRAFIK, dan DIAGRAM	vii
DAFTAR TABEL.....	viii
DAFTAR SIMBOL.....	ix
BAB 1 PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	3
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian.....	4
1.5. Manfaat Penelitian.....	4
BAB II TINJAUAN PUSTAKA.....	5
2.1. Penelitian Pendahulu	5
2.2. Pengertian Informasi	6
2.3. <i>Text Mining</i>	7
2.3.1. <i>Tokenisasi</i>	8
2.3.2. <i>Filtering</i>	9
2.3.3. <i>Stemming</i>	10
2.3.4. <i>Tagging</i>	11
2.3.5. <i>Analyzing</i>	12
2.3.5.1. Pembobotan TF-IDF.....	12
2.4. <i>Information Retrieval (IR)</i>	15
2.5. <i>Vector Space Model (VSM)</i>	20
BAB III METODOLOGI.....	23
3.1. Data Set Penelitian	23
3.2. Rancangan sistem.....	23
3.2.1. Rancangan sistem temu kembali.....	23
3.2.1.1. <i>Preprocessing</i>	25

3.2.1.2. Algoritma <i>Vector Space model</i>	36
3.3. Jadwal Kegiatan	41
BAB IV HASIL dan PEMBAHASAN	42
4.1 Pengumpulan Dokumen	42
4.1.1 Identifikasi Input	42
4.2 Analisa dan Gambaran umum sistem	45
4.2.1 <i>Preprocessing</i>	45
4.2.2 Pembobotan TF-IDF	52
4.2.3 Perhitungan <i>cosine similarity</i>	55
4.3 Uji Coba	63
4.3.1 Uji coba skenario I	64
4.3.2 Uji coba skenario II	67
4.3.3 Uji coba skenario III	69
4.3.4 Uji coba skenario IV	71
4.4 Mengukur kualitas sistem	78
BAB V KESIMPULAN dan SARAN	80
5.1. Kesimpulan	80
5.2. Saran	80
DAFTAR PUSTAKA	81
Lampiran	83

DAFTAR GAMBAR, GRAFIK, dan DIAGRAM

Gambar 2. 1 Proses <i>Text Mining</i>	8
Gambar 2. 2 Contoh Pengerjaan Proses Tokenisasi	9
Gambar 2. 3 Contoh Proses Filtering/ Stopword Removal.....	10
Gambar 2. 4 Contoh Proses Stemming	11
Gambar 2. 5 Contoh Proses Tagging	12
Gambar 2. 6 Ilustrasi Sistem Temu Kembali	15
Gambar 2. 7 Skema IR Sistem	17
Gambar 2. 8 Hubungan antar dokumen relevant dan retrieved	19
Gambar 2. 9 Ilustrasi Vector Space Model	21
Gambar 3. 1 Arsitektur Sistem Temu Kembali.....	24
Gambar 3. 2 flowchart tokenisasi.....	26
Gambar 3. 3 <i>flowchart filtering</i>	29
Gambar 3. 4 <i>flowchart stemming</i>	34

DAFTAR TABEL

Tabel 3. 1 Perhitungan TF.....	37
Tabel 3. 2 Perhitungan IDF.....	38
Tabel 3. 3 Perhitungan bobot TF-IDF.....	39
Tabel 4. 1 Tabel pertanyaan-jawaban	43
Tabel 4. 2 Nilai bobot <i>Query</i>	54
Tabel 4. 3 Panjang vector query.....	57
Tabel 4. 4 Urutan peringkat kandidat jawaban berdasarkan nilai kosinus tertinggi	60
Tabel 4. 5 Uji skenario I.....	64
Tabel 4. 6 Uji skenario II	67
Tabel 4. 7 Uji skenario III.....	69
Tabel 4. 8 Uji coba sistem <i>stemming</i> hanya pada pertanyaan.....	71

DAFTAR SIMBOL

Kode 3. 1 <i>Pseudocode</i> tokenisasi.....	27
Kode 3. 2 <i>Pseudocode filtering</i>	30
Kode 3. 3 <i>Pesudocode stemming</i>	35
Kode 4. 1 Script Tokenisasi	45
Kode 4. 2 Script Filtering.....	47
Kode 4. 3 function cek kata dasar	48
Kode 4. 4 <i>function</i> Hapus akhiran.....	49
Kode 4. 5 <i>function</i> hapus <i>derivational Prefix</i>	51
Kode 4. 6 <i>funtion stemming</i>	52
Kode 4. 7 Perhitungan bobot TF-IDF	53
Kode 4. 8 Menghitung <i>cosine similarity</i>	56