

**PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM
MENGUNAKAN JACCARD**

Tugas Akhir



Oleh :

Michel Chrisa Effendi

22 07 4256

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana

Tahun 2013

**PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM
MENGUNAKAN JACCARD**

Tugas Akhir



**Diajukan kepada Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana Sebagai salah satu syarat
dalam memperoleh gelar Sarjana Komputer**

Oleh :

Michel Chrisa Effendi

22074256

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana Yogyakarta

Tahun 2013

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM MENGUNAKAN JACCARD

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 21 Januari 2013


MICHEL CHRISA EFFENDI

22074256

HALAMAN PERSETUJUAN

Judul Skripsi : PENDETEKSIAN KEMIRIPAN DOKUMEN
BERBASIS N-GRAM MENGGUNAKAN
JACCARD

Nama Mahasiswa : MICHEL CHRISA EFFENDI

N I M : 22074256

Matakuliah : Skripsi (Tugas Akhir)

Kode : TIW276

Semester : Gasal

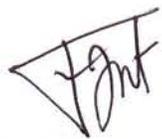
Tahun Akademik : 2012/2013

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 21 Januari 2013

Dosen Pembimbing I


Budi Susanto, SKom.,M.T.

Dosen Pembimbing II


Antonius Rachmat C., SKom.,M.Cs

HALAMAN PENGESAHAN

PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM MENGUNAKAN JACCARD

Oleh: MICHEL CHRISA EFFENDI / 22074256

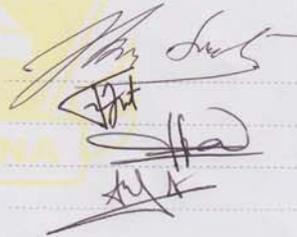
Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 11 Januari 2013

Yogyakarta, 21 Januari 2013

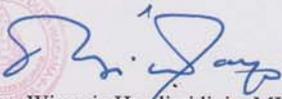
Mengesahkan,

Dewan Penguji:

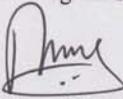
1. Budi Susanto, SKom.,M.T.
2. Antonius Rachmat C., SKom.,M.Cs
3. Yuan Lukito, S.Kom
4. Aditya Wikan Mahastama, S.Kom



Dekan


(Drs. Wimmie Hardiwidjojo, MIT.)

Ketua Program Studi


(Nugroho Agus Haryono, M.Si)

UCAPAN TERIMA KASIH

Puji syukur kepada Tuhan atas segala berkat, bimbingan, dan perlindungan-Nya sehingga penulis menyelesaikan Tugas Akhir dengan judul Implementasi Metode Jaccard dan Ngram untuk Mendeteksi Plagiasi Artikel Abstrak pada Fakultas Ekonomi Prodi Manajemen.

Penulisan laporan ini merupakan kelengkapan dan pemenuhan dari salah satu syarat dalam memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk dapat menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, sehingga dapat bermanfaat bagi penggunaannya.

Dalam menyelesaikan pembuatan program dan laporan Tugas Akhir ini, penulis telah banyak menerima bimbingan, saran dan masukan dari berbagai pihak, baik secara langsung maupun secara tidak langsung. Untuk itu dengan segala kerendahan hati, pada kesempatan ini penulis menyampaikan ucapan terimakasih kepada :

1. **Antonius Rachmat, S.Kom, M.Cs** selaku dosen pembimbing I yang telah memberikan bimbingannya dengan sabar dan baik kepada penulis, juga kepada
2. **Budi Susanto, S.Kom, M.T.** selaku dosen pembimbing II atas bimbingan, petunjuk dan masukan yang diberikan selama pengerjaan tugas ini sejak awal hingga akhir.
3. Keluarga tercinta yang selalu memberi dukungan dan do'a bagi penulis.
4. Rekan-rekan dan pihak-pihak yang tidak dapat penulis sebutkan satu persatu yang secara langsung maupun tidak langsung yang telah mendukung penyelesaian tugas ini. Terima kasih atas dukungan dan do'anya.

5. Yang terakhir dan yang terpenting, kepada Tuhan Yesus Kristus, yang memberikan seluruh kekuatan, waktu, dan kesehatan selama penulis mengerjakan Tugas Akhir ini.

Penulis menyadari bahwa laporan Tugas Akhir yang penulis susun ini masih banyak kekurangannya. Oleh karena itu, penulis mohon saran dan kritik dari pembaca guna kesempurnaan tulisan ini. Sehingga suatu saat penulis dapat memberikan karya yang lebih baik lagi.

Akhir kata penulis ingin meminta maaf bila ada kesalahan baik dalam penyusunan laporan maupun yang pernah penulis lakukan sewaktu membuat program Tugas Akhir. Sekali-lagi penulis mohon maaf yang sebesar-besarnya. Dan semoga ini dapat berguna bagi kita semua.

Yogyakarta, Juli 2012

Penulis



INTISARI

IMPLEMENTASI METODE N-GRAM DAN JACCARD UNTUK MENDETEKSI PLAGIASI ARTIKEL ABSTRAK PADA FAKULTAS EKONOMI PRODI MANAJEMEN

Plagiasi merupakan tindakan menjiplak karya seseorang dan mengakuinya sebagai karyanya sendiri. Plagiasi terhadap dokumen teks susah untuk dihindari. Oleh karena itu, sudah banyak diciptakan suatu sistem yang dapat digunakan untuk melakukan deteksi plagiasi dokumen CopyCatch, Jplag, dsb.

Untuk melakukan deteksi plagiasi dokumen pada intinya adalah dengan melakukan pencocokan *string/terms*. Dalam skripsi ini dibuat sebuah sistem yang dapat mendeteksi plagiasi pada dokumen teks dengan menggunakan metode *Jaccard* dan *Ngram*. Metode yang digunakan dalam sistem ini mempunyai fungsi yang berkesinambungan. Metode *Jaccard* digunakan untuk perangkian korpus data dan *Ngram* digunakan sebagai perhitungan *similarity*.

Hasil implementasi dari metode *Jaccard* dan *Ngram* untuk sistem pendeteksian plagiasi menunjukkan hasil persentase yang baik. Dengan penerapan metode *Jaccard* dan *Ngram* proses perhitungan persentase *similarity* menjadi lebih optimal karena akan menyisihkan 10 file teratas.

Kata kunci : Plagiarisme dokumen, *Jaccard*, *Ngram*, *stemming*, Kemiripan teks

DAFTAR ISI

HALAMAN JUDUL.....	ii
PERNYATAAN KEASLIAN SKRIPSI.....	iii
HALAMAN PERSETUJUAN.....	iv
HALAMAN PENGESAHAN.....	v
UCAPAN TERIMA KASIH.....	vi
INTISARI.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR.....	xi
DAFTAR TABEL.....	xii
BAB 1 PENDAHULUAN.....	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah.....	2
1.4 Hipotesis.....	3
1.5 Tujuan Penelitian.....	3
1.6 Metode Penulisan.....	3
1.7 Sistematika Penulisan.....	4
BAB 2 LANDASAN TEORI.....	5
2.1 Tinjauan Pustaka.....	6
2.2 Landasan Teori.....	7
2.2.1 Pengertian Kemiripan Teks.....	7
2.2.2 Pengertian Plagiarisme.....	7
2.2.3 Metode Pendeteksian Plagiarisme.....	8
2.2.4 N-gram.....	9
2.2.4.1 Pengertian N-gram.....	9
2.2.5 Jaccard.....	11
2.2.5.1 Penjelasan Jaccard Coefficient.....	11
2.2.5.2 Proses Perhitungan Similaritas Dokumen.....	12
2.2.5.3 Pengukuran Nilai Similarity.....	13
2.2.5.4 Cara Menghitung Persentase Nilai Similarity.....	15
BAB 3 ANALISIS DAN PERANCANGAN SISTEM.....	16
3.1 Kebutuhan Hardware dan Software.....	16
3.2 Spesifikasi Sistem.....	17
3.3 Arsitektur Sistem.....	19
3.4 Diagram Use Case.....	20
3.5 Perancangan Masukan.....	22
3.6 Perancangan Keluaran.....	22
3.7 Perancangan Proses.....	23

3.7.1	Preprocessing.....	23
3.7.2	Filtering dan Tokenisasi.....	26
3.7.3	Jaccard.....	27
3.7.4	N-gram.....	28
3.8	Perancangan Basis Data.....	31
3.9	Perancangan Antar Muka.....	32
3.10	Perancangan Pengujian.....	34
BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM.....		35
4.1	Antar Muka Sistem.....	35
4.1.1	Halaman Utama.....	35
4.1.2	Halaman Upload.....	35
4.1.3	Halaman Process.....	36
4.1.4	Form Tokenisasi.....	37
4.1.5	Form Perhitungan Jaccard.....	38
4.1.6	Form Perhitungan N-gram.....	38
4.2	Evaluasi Sistem.....	43
4.2.1	Pengumpulan Dokumen Artikel.....	39
4.2.2	Pra Pemrosesan Teks.....	40
4.2.3	Perhitungan Jaccard.....	40
4.2.4	Perhitungan N-gram.....	41
4.3	Analisis Sistem.....	42
4.3.1	Analisis Metode N-gram.....	42
4.3.1.1	Pengujian dengan N=5.....	44
4.3.1.2	Pengujian dengan N=6.....	44
BAB 5 KESIMPULAN DAN SARAN.....		46
5.1	Kesimpulan.....	46
5.2	Saran.....	46
DAFTAR PUSTAKA.....		47

DAFTAR GAMBAR

Gambar 2.1	Metode Pendeteksian Plagiarisme	8
Gambar 3.1	Arsitektur Sistem	19
Gambar 3.2	Use Case Diagram	20
Gambar 3.3	Flowchart Preprocessing	24
Gambar 3.4	Flowchart Get Info Dokumen	24
Gambar 3.5	Flowchart Filtering dan Tokenisasi	26
Gambar 3.6	Jaccard	27
Gambar 3.7	Parsing N-gram	28
Gambar 3.8	Perhitungan Similarity	29
Gambar 3.9	Rancangan Basis Data	31
Gambar 3.10	Upload Dokumen	32
Gambar 3.11	Halaman Perhitungan Similarity	33
Gambar 4.1	Halaman Utama	35
Gambar 4.2	Halaman Upload	36
Gambar 4.3	Halaman Proses	37
Gambar 4.4	Form Tokenisasi	38
Gambar 4.5	Form Perhitungan Jaccard	38
Gambar 4.6	Form Perhitungan N-gram	39
Gambar 4.7	Pseudocode Pra-Pemrosesan Teks	40
Gambar 4.8	Pseudocode Perhitungan Jaccard	40
Gambar 4.9	Pseudocode Perhitungan N-gram	41

DAFTAR TABEL

Tabel 2.1	Contoh Pemotongan N-gram Berbasis Karakter	10
Tabel 2.2	Contoh Pemotongan N-gram Berbasis Kata.....	10
Tabel 2.3	Contoh Pemotongan N-gram Berbasis Kata.....	11
Tabel 2.4	Perhitungan Similarity	14
Tabel 2.5	Contoh Perhitungan Similarity	14
Tabel 3.1	Arsitektur sistem.....	17
Tabel 3.2	Use Case	20
Tabel 4.1	Tabel Pengujian Dokumen Uji Plag001.txt.....	42
Tabel 4.2	Tabel Pengujian Dokumen Uji Plag002.txt.....	42
Tabel 4.3	Tabel Pengujian Dokumen Uji Plag003.txt.....	43
Tabel 4.4	Tabel Pengujian N=5	44
Tabel 4.5	Tabel Pengujian N=6	44



UKDWN

INTISARI

IMPLEMENTASI METODE N-GRAM DAN JACCARD UNTUK MENDETEKSI PLAGIASI ARTIKEL ABSTRAK PADA FAKULTAS EKONOMI PRODI MANAJEMEN

Plagiasi merupakan tindakan menjiplak karya seseorang dan mengakuinya sebagai karyanya sendiri. Plagiasi terhadap dokumen teks susah untuk dihindari. Oleh karena itu, sudah banyak diciptakan suatu sistem yang dapat digunakan untuk melakukan deteksi plagiasi dokumen CopyCatch, Jplag, dsb.

Untuk melakukan deteksi plagiasi dokumen pada intinya adalah dengan melakukan pencocokan *string/terms*. Dalam skripsi ini dibuat sebuah sistem yang dapat mendeteksi plagiasi pada dokumen teks dengan menggunakan metode *Jaccard* dan *Ngram*. Metode yang digunakan dalam sistem ini mempunyai fungsi yang berkesinambungan. Metode *Jaccard* digunakan untuk perangkungan korpus data dan *Ngram* digunakan sebagai perhitungan *similarity*.

Hasil implementasi dari metode *Jaccard* dan *Ngram* untuk sistem pendeteksian plagiasi menunjukkan hasil persentase yang baik. Dengan penerapan metode *Jaccard* dan *Ngram* proses perhitungan persentase *similarity* menjadi lebih optimal karena akan menyisihkan 10 file teratas.

Kata kunci : Plagiarisme dokumen, *Jaccard*, *Ngram*, *stemming*, Kemiripan teks

Bab 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Pada dasarnya manusia menginginkan kemudahan dalam segala hal. Sifat tersebut akan memicu tindakan negatif apabila dilatar belakangi oleh motivasi untuk berbuat curang dan rendahnya kemampuan masyarakat berkreasi dan berinovasi menciptakan suatu karya yang *original*. Dalam hal ini tindakan negatif yang dimaksud adalah plagiarism.

Fenomena kemiripan teks yang lebih spesifik sering terjadi di dunia akademis. Hal ini dikarenakan kegiatan tulis-menulis sering dilakukan oleh mahasiswa untuk menyelesaikan tugas kuliah. Praktik menduplikasikan beberapa bagian atau keseluruhan tulisan milik orang lain tanpa mencantumkan sumbernya secara teliti dan lengkap merupakan hal yang sering ditemui dalam penulisan laporan, tugas, makalah ataupun skripsi mahasiswa.

Ada dua cara untuk mengatasi permasalahan kemiripan teks, yaitu dengan mencegah dan mendeteksi. Mencegah berarti menjaga atau menghalangi agar kemiripan antar teks tidak dilakukan. Usaha seperti ini harus dilakukan sedini mungkin terutama pada system pendidikan dan moral masyarakat. Mendeteksi berarti melakukan usaha untuk menemukan tindakan plagiat yang telah dilakukan.

Banyak institusi dan tenaga pengajar menerapkan sanksi akademis terhadap pelaku kemiripan teks untuk mengurangi plagiarism. Yang menjadi permasalahannya adalah bagaimana cara untuk mengetahui apakah seorang mahasiswa melakukan plagiarism atau tidak dalam membuat suatu karya tulis. Untuk mengetahuinya perlu

dilakukan pengecekan secara teliti terhadap hasil tulisan mahasiswa tersebut, kemudian dibandingkan dengan hasil tulisan mahasiswa lainnya. Tetapi usaha tersebut akan memerlukan waktu yang lama dan ketelitian yang tinggi jika perbandingan tersebut dilakukan secara manual. Oleh karena itu diperlukan suatu sistem pendeteksian plagiarisme pada dokumen teks yang dilakukan secara terkomputerisasi.

1.2. Perumusan Masalah

Aplikasi yang akan dibuat akan mencoba menghitung berapa persen tingkat kemiripan teks dokumen uji dengan korpus data yang ada dalam database sistem. Dengan melakukan perbandingan dokumen dengan menggunakan metode *Jaccard* dan perhitungan similarity dengan menggunakan metode *N-gram* sistem akan menghasilkan nilai persentase.

Dari perkiraan tersebut, yang menjadi perumusan masalah dalam tugas akhir ini adalah sebagai berikut :

1. Apakah metode *Jaccard* dan *N-gram* dapat menghasilkan nilai persentase yang tinggi untuk sistem pendeteksian kemiripan pada dokumen abstrak yang ada pada website <http://sinta.ukdw.ac.id> yang lebih tepatnya pada prodi Ekonomi.

1.3. Batasan Masalah

Pada perhitungan tingkat plagiarisme untuk teks bersegmen pendek ini diberikan pembatasan masalah sebagai berikut :

1. Sistem akan dibangun dengan berbasis web (*web based*) dengan pengujian pada jaringan local penulis.
2. Pemilihan dokumen uji dan korpus data dilakukan secara manual oleh penulis.
3. Artikel yang digunakan sebagai data uji hanya artikel yang berbahasa Indonesia saja.

4. Proses transformasi teks meliputi penghilangan *stopword* namun tidak mencakup proses *stemming*.
5. Data yang diuji bertipe teks.

1.4. Hipotesis

Hipotesis dari penelitian ini adalah *Jaccard* dan *N-gram* dapat digunakan untuk menghitung persentase plagiasi dalam dokumen.

1.5. Tujuan Penelitian

Tujuan dari penelitian yang akan dilakukan adalah sebagai berikut:

1. Merancang aplikasi untuk mendeteksi plagiarisme dengan menggunakan *Jaccard* dan *N-gram*.
2. Mengetahui perbandingan persentase kemiripan (*similarity*) antara dokumen asli dan dokumen yang diuji dengan menggunakan algoritma *N-gram*.

1.6. Metode Penelitian

Metodologi penelitian yang dilakukan dalam tugas akhir ini dibagi dalam dua tahap yaitu sebagai berikut :

1. Tahap pengumpulan data yang akan digunakan selama pengerjaan tugas akhir ini, yang terbagi dalam beberapa langkah, yaitu :
 - a. Melakukan studi literatur mengenai konsep *information retrieval*, teori mengenai metode *Jaccard* dan *N-gram* serta pengimplementasiannya ke dalam sistem *information retrieval*.
 - b. Mencari kumpulan data atau kumpulan dokumen yang sering digunakan dalam proses sistem *information*.

2. Tahap pengembangan sistem.

Proses yang terjadi dalam pengembangan sistem terdiri dari 2 tahap, yaitu :

a. Pra pemrosesan

Tahap pra pemrosesan dilakukan proses tokenisasi, penghilangan stopword dan pencatatan teks unik.

b. Pemrosesan

Tahap pemrosesan dilakukan dengan menerapkan metode *Jaccard* dan *N-gram* untuk perhitungan persentase dokumen yang dibandingkan.

1.7. Sistematika Penulisan

Penulisan laporan tugas akhir ini akan dibagi menjadi 5 (lima) bagian, yaitu :

Bab 1 memberikan gambaran umum mengenai hal yang akan diteliti oleh penulis dalam tugas akhir ini. Pendahuluan memuat latarbelakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metode penelitian, dan sistematika penulisan laporan.

Bab 2 ini terdiri dari dua sub bab, yaitu tinjauan pustaka dan landasan teori. Tinjauan pustaka memaparkan penelitian-penelitian terdahulu beserta teori yang berkaitan dengan topik penelitian yang diambil oleh penulis, sedangkan landasan teori berisi konsep-konsep yang digunakan dalam mendukung penelitian ini.

Bab 3 terdiri dari beberapa sub bab yang digunakan dalam perancangan sistem, antara lain spesifikasi sistem, arsitektur sistem, diagram *use case*, algoritma dalam membangun sistem, rancangan antarmuka sistem, dan rancangan pengujian sistem.

Bab 4 ini membahas mengenai implementasi serta pengujian sistem yang telah dibangun oleh penulis berdasarkan pada rancangan sistem yang telah diuraikan pada Bab3. Bab ini juga berisi hasil dari proses yang dilakukan oleh sistem dan analisis dari sistem yang telah berjalan.

Bab 5 ini berisi kesimpulan dari penelitian yang dilakukan oleh penulis beserta saran yang diberikan oleh penulis bagi penelitian- penelitian mendatang yang memiliki topik yang sama dengan penelitian ini.

© UKDW

BAB 5

Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil implementasi dan analisis sistem yang menggunakan metode Jaccard dan Ngram, maka didapat kesimpulan sebagai berikut:

1. Penggunaan metode *Jaccard* dan *Ngram* dapat mendeteksi tingkat kemiripan dokumen yang berbeda sehingga dapat diketahui apakah satu dokumen merupakan plagiasi atau tidak.
2. Penggunaan metode *Ngram* untuk perhitungan nilai similarity antara dokumen uji dengan 10 korpus data yang memiliki *index Jaccard* terbesar.

5.2 Saran

Beberapa saran yang dianjurkan untuk pengembangan dan perbaikan sistem adalah sebagai berikut :

1. Dokumen uji dapat diambil langsung dari suatu alamat URL tanpa harus melakukan upload dokumen terlebih dahulu.
2. Menggunakan dukungan bahasa pemrograman *AJAX* untuk menghasilkan antar muka yang lebih interaktif.

DAFTAR PUSTAKA

KBBI, 1997: 775

Even-Zohar, Y. 2002. *Introduction to Text Mining*, Supercomputing.

Kosinov, Serhiy. 2002. *Evaluation of N-Grams Conflation Approach in Text-Based Information Retrieval*. University of Alberta. Canada

Stein, B., Meyer, S. zu Eissen. 2006. *Near Similarity Search and Plagiarism Analysis*, 29th Annual Conference of the German Classification Society(GfKI), Magdeburg, ISDN 1431-8814, pp. 430 – 437.

Leydesdorff, L. (2008). *On the Normalization of Author Co-Citation Data: Salton's Cosine versus the Jaccard Index*. Journal of the American Society of Information Science & technology. New York: Wiley & Sons Inc. Vol 59, 77-85.

Iyer, Parvati., Singh, Abhipsita. 2005. *Document Similarity Analysis for a Plagiarism Detection System*, 2nd Indian International Conference on Artificial Intelligence (IICAI-05), pp. 2534 – 2544.

Mutiara, Benny; Agustina, Sinta. 2008. *Anti Plagiarsm Application with Algorithm Karp-Rabin at Thesis in Gunadarma University*. Gunadarma University. Depok, Indonesia

Ridhatillah, Ardini . 2003. *Dealing with Plagiarism in the Information System Research Community: A Look at Factors*

that Drive Plagiarism and Ways to Address Them,

Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Tan, A. 1999. *Text Mining: The state of the art and the challenges*,
In Proc of the Pacific Asia Conf on Knowledge Discovery and
Data Mining PAKDD'99 workshop on Knowledge Discovery
from Advanced Databases.

© UKDWN