PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM MENGGUNAKAN 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 MENGGUNAKAN 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 MENGGUNAKAN 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	Dosen Pembimbing II
Budi Susanto, SKom.,M.T.	Antonius Rachmat C., SKom.,M.Cs

HALAMAN PENGESAHAN

PENDETEKSIAN KEMIRIPAN DOKUMEN BERBASIS N-GRAM MENGGUNAKAN 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 Handiwidjojo, 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 penggunanya.

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 teks 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 perangkingan 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

DAFTAR GAMBAR

DAFTAR TABEL

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 teks 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 perangkingan 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 pembandingan tersebut dilakukan secara manual. Oleh karena itu diperlukan suatu sistem pendeteksian plagiarism 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 perangkingan 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 plagiasi 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 plagiarism 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 literature 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 pengujians istem 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.

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 PlagiarismAnalysis*, 29th Annual Conference of the German Classification Society(GfKl), 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 Intelegence (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., & Schűtze, 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.