DETEKSI PLAGIASI OTOMATIS BERBASIS N-GRAM

Tugas Akhir

oleh
ANDREAN CANDRA WIJAYA
22084599

Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Tahun 2012

DETEKSI PLAGIASI OTOMATIS BERBASIS N-GRAM

Tugas Akhir

Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

ANDREAN CANDRA WIJAYA
22084599

Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Tahun 2012

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

Deteksi Plagiasi Otomatis Berbasis N-Gram

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, 1 Juli 2012

ANDREAN CANDRA WIJAYA
22084599

HALAMAN PERSETUJUAN

Judul Skripsi	: Deteksi Plagiasi Otomatis Berbasis N-Gram
Judul	: ANDREAN CANDRA WIJAYA
N I M	: 22084599
Matakuliah	: Tugas Akhir
Kode	: TIW276
Semester	: Genap
Tahun Akademik	: 2011/2012

Telah diperiksa dan disetujui di Yogyakarta,
Pada tanggal 1 Juli 2012

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, M.A.	Dra. Widi Hapsari, M.T.

HALAMAN PENGESAHAN

DETEKSI PLAGIASI OTOMATIS BERBASIS N-GRAM

Oleh: ANDREAN CANDRA WIJAYA / 22084599

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
20 Juni 2012

Yogyakarta, 1 Juli 2012
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, M.A.
2. Dra. Widi Hapsari, M.T.
3. Ir. Sri Suwarno, M.Eng.
4. Budi Susanto, SKom.,M.T.

Dekan

(Drs. Wimmie Handiwidjojo, MIT.)

Ketua Program Studi

(Nugroho Agus Haryono, M.Si)

INTISARI

Deteksi Plagiasi Otomatis Berbasis N-Gram

Untuk mengetahui bahwa dokumen termasuk dalam kategori plagiasi biasanya seseorang membandingkan isi antara dokumen yang akan diujikan (*suspicious document*) dengan dokumen dari sumber aslinya. Akan tetapi ini dapat menyebabkan seseorang merasa kewalahan apabila ada banyak dokumen yang akan dibandingkan.

Untuk menangani masalah tersebut maka diperlukan pendeteksian plagiasi otomatis yang mampu menampilkan dokumen sumber plagiasi beserta bagian kalimat yang diplagiasi. Pada penelitian ini akan digunakan metode menggunakan N-gram dengan N=6 dalam melakukan pendeteksian, serta perangkingan dokumen sumber dengan menggunakan tf-idf normalisasi dan DICE *coefficient.*

Hasil dari penelitian ini, pendeteksian plagiasi dengan menggunakan N-Gram tidak selalu dapat mendeteksi bagian kalimat dengan baik dikarenakan faktor perangkingan yang kurang akurat serta terdapat bagian kalimat yang tidak terdeteksi apabila kata kurang dari enam.

UCAPAN TERIMA KASIH

Puji dan syukur kehadirat Tuhan Yang Maha Esa, yang telah melimpahkan rahmat dan karunia kepada penulis dapat menyelesaikan skripsi dengan judul Deteksi Plagiasi Otomatis Berbasis N-Gram dengan baik dan tepat waktu.

Penulisan dan penyusunan skripsi ini disusun dalam rangka melengkapi syarat untuk memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, dan dapat bermanfaat bagi penggunanya

Pada kesempatan ini, penulis mengucapkan terima kasih kepada semua pihak yang telah membantu dalam menyusun skripsi, antara lain :

1. **Bu Lucia Dwi Krisnawati, M.A.** selaku dosen pembimbing I yang telah memberikan bimbingannya dengan baik dan sabar, juga kepada
2. **Bu Widi Hapsari, Dra., M.T.** selaku dosen pembimbing II yang memberikan petunjuk dan masukkan dari awal hingga akhir selesainya skripsi ini.
3. Keluarga tercinta yang selalu memberikan semangat, perhatian, dan motivasi agar skripsi ini selesai.
4. Voni, Riky, David, dan pihak lain yang tidak dapat penulis sebut satu-persatu yang telah memberikan semangat dan masukan, sehingga skripsi ini dapat terselesaikan dengan baik dan tepat waktu.

Akhir kata, dengan kerendahan hati, penulis menyadari bahwa skripsi ini masih jauh dari sempurna, oleh karena itu penulis menerima kritik, saran, semoga skripsi ini dapat bermanfaat bagi semua pihak.

Yogyakarta, Mei 2012

Andrean Candra Wijaya

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Kemajuan teknologi mempermudah para pelajar dalam mengakses informasi, terutama yang berhubungan dengan akademik. Kemudahan dalam mengakses informasi ini memberikan peluang bagi seorang pelajar untuk melakukan tindak plagiasi dari suatu karya intelektual milik orang lain. Seorang pelajar dapat dikatakan melakukan tindakan plagiasi apabila suatu kalimat karya milik orang lain diubah susunan letak katanya atau melakukan pengambilan kalimat secara identik tanpa melakukan parafrase.

Dalam melakukan pendeteksian plagiasi biasanya seseorang membandingkan isi antara dokumen yang akan diujikan (*suspicious document*) dengan dokumen dari sumber aslinya. Akan tetapi ini dapat menyebabkan seseorang merasa kewalahan apabila ada banyak dokumen yang akan dibandingkan, maka dari itu diperlukan pendeteksian plagiasi secara otomatis.

Dalam penelitian ini akan dilakukan percobaan atau eksperimen dalam melakukan pendeteksian plagiasi dan menampilkannya dengan menggunakan tf-idf yang di normalisasi dan DICE *coefficient* yang akan digunakan sebagai *filtering*, serta secara *exhaustive search* akan dilakukan pendeteksian dengan menggunakan n-gram.

1.2. Perumusan Masalah

Permasalahan pada penelitian ini yaitu:

1. Bagaimana cara menemukan segmen yang dianggap plagiasi.
2. Bagaimana cara menemukan segmen yang dianggap plagiasi pada dokumen sumber

3. Bagaimana cara menampilkan segmen yang dianggap plagiasi pada dokumen uji atau query (*suspicious document*) dan pada dokumen sumber (dokumen korpus).

1.3. Batasan Masalah

1. Query berupa dokumen dengan sekitar kata atau satu paragraph tidak termasuk gambar.
2. Dokumen korpus diambil dari abstrak karya mahasiswa UKDW yang terdapat pada situs SINTA.
3. Sistem akan di buat dan diimplementasikan berbasiskan web dan di buat dengan inputan dokumen dalam bentuk file teks (*.txt).
4. Pendeteksian yang dilakukan tidak memperhatikan struktur dari kata maupun makna dari kalimat tersebut (hanya *copy paste*).
5. Pendeteksian yang dilakukan berdasarkan kata bukan karakter.
6. Sistem hanya mampu menghasilkan 15 dokumen berdasarkan hasil dari perangkingan. Hasil dari perangkingan dokumen ini akan digunakan untuk pendeteksian.

1.4. Tujuan Penelitian

1. Implementasi n-gram dalam melakukan deteksi plagiasi otomatis dengan tf-idf normalisasi dan DICE *coefficient.*
2. Melihat efektifitas dari penerapan metode n-gram dalam deteksi plagiasi otomatis dengan filtering.

1.5. Metode / Pendekatan

Metode yang akan dilakukan dalam penelitian adalah sebagai berikut:

1. Melakukan studi pustaka dengan cara mencari informasi-informasi atau literatur yang berhubungan dengan penelitian,

2. Melakukan analisa terhadap informasi dan literatur yang dipelajari.
3. Mengimplementasikan metode n-gram dalam melakukan pendeteksian dengan menggunakan tf-idf normalisasi dan DICE *coefficient* dalam melakukan perangkingan dokumen.

1.6. Sistematika Penulisan

Penulisan laporan skripsi dibagi menjadi lima bab dengan sistematika sebagai berikut:

BAB 1 PENDAHULUAN yang berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, metode/pendekatan, dan sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA yang berisi tinjauan pustaka dan landasan teori yang akan digunakan pada penelitian.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM yang berisi rancangan sistem yang akan dibuat berdasarkan informasi dan literatur yang sudah dipelajari.

BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM yang berisi hasil dari penelitian/implementasi yang dibuat.

BAB 5 KESIMPULAN DAN SARAN yang berisi pernyataan singkat dari sistem yang dibuat, serta saran untuk kegiatan penelitian sehingga lebih baik lagi.

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan bahwa tidak semua bagian kesamaan kalimat dideteksi oleh sistem dengan baik, pendeteksian bagian kalimat yang dapat dideteksi oleh sistem adalah bagian kalimat yang memiliki kata minimal enam. Selain itu, dari hasil percobaan ditemukan bahwa pendeteksian plagiasi dengan menggunakan perangkingan menjadi kurang efektif. Hal ini disebabkan sistem tidak mampu melakukan pengambilan (perangkingan) dokumen yang dikarenakan bagian kalimat yang plagiasi berupa kata-kata yang sering muncul di dokumen sumber, sehingga bagian kalimat yang seharusnya terdeteksi sebagai plagiasi menjadi tidak terdeteksi. Pendeteksian plagiasi pada sistem ini menghasilkan nilai presisi dengan rata-rata 0.95 dan nilai recall dengan rata-rata 0.91934226754293, dengan catatan pendeteksian plagiasi yang dibuat tidak menghiraukan bahwa dokumen plagiasi mencantumkan asal dokumen sumber atau tidak.

5.2. Saran

Sistem yang digunakan merupakan sistem pendeteksi plagiasi eksternal (*copy-paste*), maka dari itu pengembangan sistem untuk kedepannya adalah sistem tidak hanya terbatas dalam mendeteksi *copy-paste* saja, melainkan mampu mendeteksi dan menampilkan kata-kata atau bagian kalimat yang telah dimodifikasi (plagiasi intrinsik).

DAFTAR PUSTAKA

Barron-Cedeno, A., & Rosso, P. (2009). On Automatic Plagiarism Detection Based on n-Grams Comparison. *Proceeding ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (pp. 696-700). Heidelberg: Springer-Verlag Berlin.

Barron-Cedeno, A., & Rosso, P. (2010). Towards the 2nd International Competition on Plagiarism Detection and Beyond*. *Proc. 4th Plagiarism Int. Conf.* Newcastle.

Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES*.

Huang, D. H., & Yang, M. Y. (2011). Retrieved February 12, 2012, from School of Information Technology and Electrical Engineering University of Queensland: http://itee.uq.edu.au/%7Einfs4203/Lecture/Lesson07_Text_Mining_2011.pdf

Potthast, M., Stein, B., Eiselt, A., Barr'on-Cedeno, A., & Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection*.

Wibowo, A. (2011). Pengujian Kerelevanan Sistem Temu Kembali Informasi.

Zechner, M., Muhr, M., Kern, R., & Granitzer, M. (2009). External and Intrinsic Plagiarism Detection Using Vector Space Models. *SEPLN conference*, (pp. 47-55).