SISTEM DETEKSI KEMIRIPAN ANTAR DOKUMEN TEKS MENGGUNAKAN ALGORITMA SCAM

Hevi Oktaviani, 12 1065 1056 Program Studi Teknik Informatika

Abstrak

SCAM adalah singkatan dari *Stanford Copy Analysis Mechanism* dimana menjadi biasanya tolak ukur relatif untuk mendeteksi *overlapping* dengan membuat perbandingan pada satu set kata-kata yang umum antara dokumen tes dan dokumen terdaftar.

Vector Space Model sebuah model populer dalam domain IR, adalah model VSNI. Mengingat permintaan dengan bobot yang sesuai, produk dot terjadinya vektor tertimbang dari query dengan dokumen yang disimpan dan dihitung: jika nilai produk dot melebihi batas tertentu, dokumen ditandai untuk mencocokkan query.

Sistem ini mampu melakukan mengukuran kemiripan dokumen Bahasa Indonesia. Dokumen yang digunakan berupa file import dokumen dan input text jadi user dapat memilih input data dengan.

Kata Kunci: Scam, overlapping, Vector Space Model.

1. Latar Belakang

Digitalisasi pengolahan informasi dengan menggunakan komputer menghasilkan fasilitas yang copy-paste (salin-tempel) sehingga memudahkan pengolahan data sesuai dengan kebutuhan misalnya memenuhi tugas kuliah, membuat paper, dan sebagainya. Hal ini tentu berpotensi terjadinya tindakan penjiplakan suatu karya tulis tanpa ijin seperti plagiat. Selain itu, semakin meningkat kecanggihan manusia dalam melakukan plagiat juga menjadi alasan lain dibutuhkannya sistem yang dapat membantu mendeteksi plagiat dalam sebuah dokumen. Oleh karena itu, penulis menggagas untuk dibangunnya sebuah sistem deteksi plagiat (Plagiarism detection system) yang menggunakan (Stanford **SCAM** Copy Analysis Mechanism) untuk melakukan pencocokan kemiripan antar dokumen dengan menyajikan dokumen plagiat yang disusun berdasarkan kemiripan dokumen. SCAM sendiri mempunyai ukuran relatif yang dapat mendeteksi tumpang tindih (overlap). Dengan adanya sistem deteksi plagiat ini diharapkan akan membantu proses deteksi plagiat.

2. Penerapan Pengukuran Kesamaan Dokumen

Pengukuran kemiripan dokumen dapat digunakan untuk mengukur kemiripan dokumen resmi, dokumen standar seperti Standar Operasional Procedure (SOP), peraturan perundangan, hasil penelitian dan lain-lain. Pengukuran dokumen dapat diterapkan di berbagai instansi seperti pemerintahan, perusahaan, akademik dan lain-lain. Plagiarisme dalam bidang akademik dapat dibagi menjadi dua (Vamplew & Dermaoudly, 2015) yaitu:

- a. Content-based file comparison
 Pendekatan content-based file
 merupakan pendekatan untuk
 mengukur kemiripan dokumen
 atau teks seperti tugas essay
 mahasiswa dan hasil penelitian.
- b. Content-based comparison of source code Pendekatan ini digunakan untuk mengukur kemiripan source code pemrograman.

3. Pengukuran Kemiripan

Pada sub bab ini dijelaskan tentang jenis pengukuran kemiripan, pendekatan pengukuran kemiripan, penelitian-penelitian yang telah dilakukan terhadap pendekatan pengukuran kemiripan.

4. Algoritma SCAM

SCAM adalah singkatan dari Stanford Copy Analysis Mechanism dimana menjadi biasanya tolak ukur relatif untuk mendeteksi overlapping dengan membuat perbandingan pada satu set kata-kata yang

umum antara dokumen tes dan dokumen terdaftar.

Secara general algoritma ini bekerja dengan memberikan pendekatanpendekatan khusus untuk mendeteksi kesamaan antara dokumen. Dan langkahlangkah deteksi tersebut, dibagi menjadi empat langkah utama, yaitu:

- 1. Pengindeksan *dataset*: Dataset ini terdiri dari beberapa dokumen yang preproses dan di indeks.
- 2. Pengolahan dokumen uji: Dokumen yang akan diuji diberikan dalam input diproses untuk proses *tokenize* yang akan dimasukkan dalam daftar kata yang akan diolah nantinya.
- 3. Pencarian pada indeks : Indeks akan digunakan untuk mengambil kecocokan antara dokumen uji dan dokumen milik *dataset*.
- 4. Mengevaluasi kesamaan: Menggunakan rumus SCAM. perbandingan kesamaan dokumen, nantinya akan berdasarkan dokumen uji dan dokumen milik *dataset*.

5. Implementasi Sistem

Berdasarkan hasil analisis dan perancangan sistem yang telah dilakukan, maka dilakukan implementasi sistem deteksi kemiripan dokumen dengan metode *SCAM* ke dalam bentuk program dengan menggunakan bahasa pemrograman PHP dan database MySQL.

5.1 Tampilan Utama Sistem

Tampilan utama merupakan halaman yang akan dijalankan ketika sistem dijalankan, halaman utama menampilkan nama aplikasi dan penyusun penelitian, halaman utama dapat dilihat seperti gambar dibawah ini:



Gambar 4.1. Halaman Utama

5.2 Halaman Data Set

Pada tampilan menu Data Set, akan menampilkan sebuah tabel dimana Pengguna dapat memilih berapa jumlah dataset yang digunakan, pada halaman ini user disediakan beberapa fitur yaitu Tambah, Edit dan Hapus. Tampilan halaman dataset dapat kita lihat pada gambar 4.2.



Gambar 4.2. Halaman Dataset

Pada gambar 4.2 ada beberapa fitur yaitu cari dokumen digunakan untuk mencari dokumen dataset, tombol tambah digunakan untuk membuka halaman form tambah dokumen dataset seperti gambar 4.3, tombol edit merupakan fitur untuk melakukan perubahan dataset dan jika user memilih hapus maka user akan menghapus dataset.



Gambar 4.3. Halaman Form Tambah Dokumen Dataset

Nama Dolumen Nama Dolumen ALASTRIANS TENS CENIAAN NAVE SEES CLASSFER (NBC) (NTILK PENALCLAMPOWN TENS SERITA. DAN ASSTRUCT ANACEMIS ALASTRIANS TENS CENIAAN NAVE SEES CLASSFER (NBC) (NTILK PENALCLAMPOWN TENS SERITA. DAN ASSTRUCT ANACEMIS Tingaps a Vitam columen tens in rejuct sells allations and bridges author best dan allations ackering dies logical dans bestell solls and bestellage under best dan allations ackering dereitand dans members dans legislations bestellage in der bestellage in de	ormulir Tambah	atau Edit Dokumen Data Set
Abmade/Herinangan informasities diguid teith introdu sargari orgat. Soal in dipersistant 50% leis diguid diata herinant data terchinata filinggan valume diamine heis in dipersista indicata on telesgan sarber diata antibilita sabardia perjemania limiti yang malai menengati. Aebituhan anadas leit mening sangdi dipersistan diata mengana leis sang pitak terchinata terberuh. Salah sahi selepan perdipi diata herinanga padah likadiksia salah selepansa seles salah garinan seles sedan salah menilak berbaga can perdekatan antibia lein perdikksian probabilidiri. Sappori verbor mantine, dian antibial resar selevani, salah depision berbaga can perdekatan antibia lein perdikksian probabilidiri. Sappori verbor mantine, dan antibial resar selevani, salah depision berbaga can perdekatan antibial seminanti selestan (PRC) menilal selestanga selebah kelebahanna diama kempatangan, Sahum metode in melilik selematan datah sumuni saga sala dipersisi, salah kelepandera filias selestan selestanda datah sumuni saga sala dipersisi selestan menganakan datah dibuman berbagai dibuturne selestan se	Nama Dokumen	Nasifikasi teks dengan nalve bayes
	isi Dokumen	Abstacl/elembangs in thromas less digital from humbs angun cope. Sall an dispensional 90% less digital datum humbs and lost bestitution. Riggins values dominier less in deput an infession of brediga some breds data altibitati some date legislation in the cytoriesm portional humbs grant an meraligia. Reductional audios lest mines grande dispension datum mensanges less supplicate less and production. Sall situal situation productions may be a lost productions and situation sold in mensanges less supplicate less and berdaga can predictation audios in predictation production. Supplication composition in the legislation and in mensional berdaga can predictation audios in predictation production. Supplication composition is supplicated in the legislation and productions. Supplication is supplicated in the legislation and supplications and supplication is supplicated in the legislation and supplications. Supplication is supplicated in the legislation and supplications and supplications and supplications. In the legislation and supplications are supplicated in the legislation and supplications and supplications. All supplications are supplicated and supplications and supplications and supplications. All supplications are supplicated and supplications and supplications and supplications. All supplications are supplications and

Gambar 4.4 Halaman form edit data set

5.3 Halaman Testing

Menu halaman testing akan menampilkan sebuah form input berupa teks dokumen uji, dimana sistem akan menggunakan input text atau import dokumen untuk melakukan testing kesamaan dokumen sistem akan mengolah data yang masuk dari user tersebut dan kemudian diuji apa dokumen memiliki kemiripan dengan dataset.



Gambar 4.5. Halaman Form Input Dokumen Uji



Gambar 4.6. Halaman Hasil Uji

6. Kesimpulan

Berdasarkan analisis dan pengujian yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut :

- Aplikasi ini mampu melakukan mengukuran kemiripan dokumen Bahasa Indonesia.
- Dokumen yang digunakan berupa file import dokumen dan input text jadi user dapat memilih input data dengan du acara.

7. Saran

Dalam pengembangan penelitian ini penulis memberikan beberapa saran, yaitu:

- 1. Untuk pengembang berikutnya sistem dapat ditambahkan fitur stemmer untuk memperoleh akurasi yang lebih efektif.
- Algoritma pendeteksian kesamaan dokumen ini dapat dikembang dengan membandingkan algoritma yang lain seperti Rabin Karp, Jaro-Winkler.

DAFTAR PUSTAKA

Agustina, S., 2008, Aplikasi Anti Plagiatisme Dengan Algoritma Karp-

Rabin Pada Penulisan Ilmiah Universitas Gunadarma, Jakarta. Al-Bahra bin Ladjamudin. 2005. Analisis dan Desain Sistem Informasi.

Yogyakarta: Graha Ilmu.

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., dan Moeliono, M., 1998, Tata

Bahasa Baku Bahasa Indonesia, Departemen Pendidikan dan Kebudayaan Republik Indonesia, Jakarta.

Arief, M.Rudianto. 2011. Pemrograman Web Dinamis Menggunakan Php dan Mysql. Yogyakarta: ANDI.

Elbegbayan, N., 2005, Winnowing, a Document Fingerprinting Algorithm,

TDDC03 Projects, Spring.

Firdaus, H. B., 2003, Deteksi Plagiat Dokumen Menggunakan Algoritma

RabinKarp, Jurnal Ilmu Komputer Dan Teknologi Informasi.

Jogiyanto. 2005. Analisis dan Desain Sistem Informasi. Yogyakarta: Penerbit Andi.

Kementerian Pendidikan Nasional, 2010, Peraturan Menteri Nasional

> Republik Indonesia Nomor 17 Tahun 2010 Tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi, Jakarta.

Novanta, A., 2009, Pendeteksian Plagiarisme pada Dokumen Teks dengan Menggunakan Algoritma Smith-Waterman, Medan.

Nugroho, Adi. 2006. E-commerce. Informatika Bandung. Bandung.

Schleimer, S., Wilkerson, D. S., and Aiken, A., 2003, Winnowing: Local

Algorithms for Document Fingerprinting, SIGMOD, San Diego California.

Soeherman, Bonnie & Pinontoan, Marion. 2008. Designing Information

System. Jakarta : Elex Media Komputindo.

Vamplew, P., and Dermaoudly, J., 2005, An Anti-Plagiarism Editor for

Software Development Courses, Conferences in Research and Practise in information Technology, Vol 42, Australia.