DETEKSI PLAGIARISME DOKUMEN BAHASA INDONESIA DENGAN ALGORITMA JARO-WINKLER *DISTANCE*

YUDHA ANGGARA, 12 1065 1028
PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS MUHAMMADIYAH JEMBER, 2016

Abstrak

Perkembangan pendidikan yang semakin berkembang pesat membuat proses pembuatan karya tulis semakin mudah dan cepat. Penulisan karya ilmiah yang dibuat tidak menutup kemungkinan terdapat penulisan yang sama, untuk dapat mengetahui tingkat kesamaan dokumen teks dengan cepat maka perlu alat bantu untuk menghitung tingkat kesamaan antar dokumen. Pada penelitian ini akan dibuat sebuah aplikasi untuk menghitung tingkat kesamaan dokumen teks berbahasa Indonesia berbasis web, dengan menerapkan algoritma *Jaro-Winkler distance*. Tujuan dari penerapan algoritma ini adalah membandingkan kesamaan antar dokumen teks berbahasa Indonesia, sehingga dapat ditentukan sebuah dokumen tersebut plagiat atau tidak. Aplikasi ini dirancang menggunakan bahasa pemrograman PHP dan database MySQL. Pengujian terhadap aplikasi menggunakan data abstrak jurnal skripsi program studi Teknik Informatika Universitas Muhammadiyah Jember berbahasa Indonesia yang berjumlah 100 buah paper. Dari hasil analisis dokumen uji 1 memiliki kesamaan tertinggi dengan dokumen nomer 55 dengan nilai 86,267%, dokumen 2 memiliki kesamaan tertinggi dengan dokumen 66 dengan nilai 95,922% dan dokumen 3 memiliki kesamaan tertinggi dengan dokumen 23 dengan nilai 98.361 %.

Kata Kunci : Algoritma Jaro-Winkler distance, PHP, MySQL, Deteksi Kesamaan Dokumen.

1. Latar Belakang

Dengan semakin berkembangnya teknologi informasi, sehingga membuat pembuatan karya tulis semakin mudah dan cepat. Hal tersebut dikarenakan informasi kini tersedia secara melimpah. Akan tetapi dikarenakan kemudahan dalam memperoleh informasi tersebut, pada pembuatan karya tulis sering ditemukan kesamaan dengan karya tulis orang lain sehingga kemudian menimbukan isu plagiarisme.

Menurut *Tia Septiani Widi* (2012), plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain.

Jaro-Winklrer distance merupakan varian dari *Jaro distance metric* yang merupakan sebuah algoritma untuk mengukur kesamaan antara dua string, biasanya algoritma ini digunakan

di dalam pendeteksian duplikat dokumen. Penelitian ini akan membahas mengenai pendeteksian plagiarisme dari sebuah dokumen dengan melakukan perbandingan dengan dokumen lainnya yang memanfaatkan metode pencocokan string pada dokumen.

Berdasarkan masalah yang telah dikemukakan, maka peneliti mengambil penelitian Tugas Akhir dengan Judul **"Deteksi Plagiarisme Dokumen Bahasa Indonesia Dengan Algoritma Jaro-Winkler Distance"** dengan harapan mampu mengatasi kendala yang telah dikemukakan tersebut.

2. Algoritma

Pemrograman sudah menjadi kegiatan yang sangat penting di era teknologi informasi saat ini. Program yang berjalan diberbagai device seperti komputer (personal computer), netbook, handheld, web (berbasis internet) pada dasarnya tidak dibangun begitu saja, melainkan ada suatu proses yang menjadi suatu pola kerja dari program itu sendiri yaitu algoritma.

Algoritma dalam pengertian modern mempunyai kemiripan dengan istilah resep, proses, metode, teknik, prosedur, rutin,. Algoritma adalah sekumpulan aturan-aturan berhingga yang memberikan sederetan operasi-operasi untuk menyelesaikan suatu jenis masalah yang khusus (Knuth, 1973). Berdasarkan pengertian algoritma diatas, dapat disimpukan bahwa algoritma merupakan suatu istilah yang luas, yang tidak hanya berkaitan dengan dunia komputer.

Kriteria Algoritma (Knuth, 1973) adalah:

1. Input: algoritma dapat memiliki nol atau lebih masukan dari luar.
2. Output: algoritma harus memiliki minimal satu buah hasil keluaran.
3. Definiteness (pasti): algoritma memiliki instruksi-instruksi yang jelas dan tidak ambigu.
4. Finiteness (ada batas): algoritma harus memiliki titik berhenti (stopping role).
5. Effectiveness (tepat dan efisien): algoritma sebisa mungkin harus dapat dilaksanakan dan efektif.

2.1.Algoritma Jaro-Winkler Distance

Jaro-Winkler Distance adalah sebuah algoritma untuk mengukur kesamaan antara dua string, biasanya algoritma inni digunakan dalam pendeteksian duplikat. Semakin tinggi *Jaro-Winkler distance* untuk dua string, semakin mirip dengan string tersebut. *Jaro-Winkler* distance terbaik dan cocok untuk digunakan dalam perbandingan string singkat seperti nama orang. Skor normalnya seperti 0 menandakan tidak ada kesamaan, dan 1 adalah sama persis.

Dasar dari algoritma ini memiliki tiga bagian:

1. Menghitung panjang string
2. Menghitung jumlah karakter yang sama di dalam dua string
3. Menemukan jumlah *transposisi*

Pada algoritma Jaro digunakan rumus untuk menghitung jarak (*dj*) antara dua string yaitu s1 dan s2.

$$d_j = \frac{1}{3} x \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{|m|} \right) \dots\dots\dots (1)$$

dimana :

m = jumlah karakter yang sama persis

|s1| = panjang string 1

|s2| = panjang string 2

t = jumlah tranposisi

jarak teoritis dua buah buah karakter yang disamakan dapat dibenarkan jika tidak melebihi:

$$\left(\frac{max(|s_1|, |s_2|)}{s} \right) < -1 \dots\dots\dots (2)$$

Akan tetapi bila mengacu kepada nilai yang akan dihasilkan oleh algoritma *Jaro-Winkler* maka nilai jarak maksimalnya adalah 1 yang menandakan kesamaan string yang dibandingkan mencapai seratus persen atau sama persis. Biasanya s1 digunakan sebagai acuan untuk urutan di dalam mencari transposisi. Yang dimaksud transposisi di sini adalah karakter yang sama dari string yang dibandingkan akan tetapi tertukar urutannya.

Sebagai contoh, dalam membandingkan kata CRATE denganTRACE, bila dilihat seksama maka dapat dikatakan semua karakter yang ada di s1ada dan sama dengan karakter yang ada di s2, tetapi dengan urutan yang berbeda. Dengan mengganti C dan T, dapat dilihatperubahan kata CRATE menjadi TRACE. Pertukaran dua elemen string inilah adalah contoh nyata dari transposisi yang dijelaskan. Dalam pencocokkan DwAyNE dan DuANE memiliki urutan yang

sama D-A-N-E, jadi tidak ada transposisi.

Jaro-Winkler distance menggunakan prefix scale(p) yang memberikan tingkat penilaian yang lebih, dan prefix length (l) yang menyatakan panjang awalan yaitu panjang karakter yang sama dari stringyang dibandingkan sampai ditemukannya ketidaksamaan. Bila string s1 dan s2 yang diperbandingkan, maka Jaro-Winkler distancenya (dw) adalah :

$$d_w = \quad d_j + \left(lp\left(1 - d_j\right)\right) \ldots\ldots (3)$$

dimana :

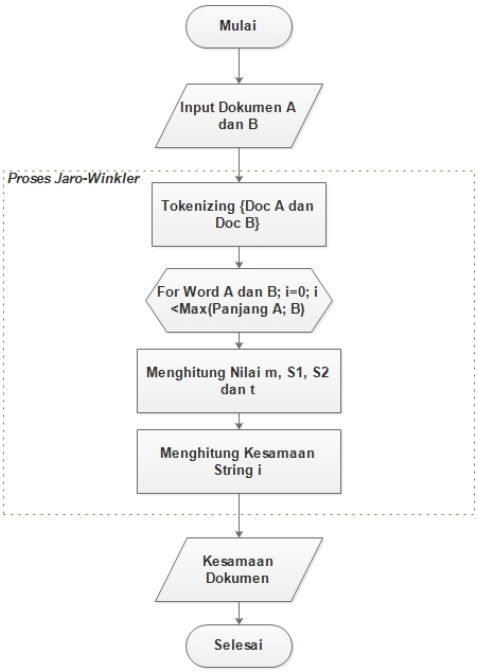
dj = Jaro distance untuk string s_1 dan s_2

l = panjang pefiks umum di awal string nilai maksimalnya 4 karakter (panjang karakter yang sama sebelum ditemukan ketidak samaan max 4)

p = konstanta *scaling factor*. Nilai standar untuk konstanta ini menurut Winkler adalah p = 0,1.

3. Flowchart Sistem

Flowchart sistem atau diagram alir merupakan teknik untuk menggambarkan logika prosedural atau jalur kerja sistem yang akan dibangun. Berikut ini adalah diagram alir untuk proses deteksi plagiarisme dokumen Bahasa Indonesia menggunkan algoritma Jaro-Winkler.



Gambar 3.1. Flowchart Sistem

4. Implementasi Sistem

4.1.Tampilan Aplikasi

Setelah semua persiapan teknis dilakukan, selanjutnya menjalankan aplikasi deteksi plagiasi dengan menggunakan algoritma *jaro-winkler distance.*

4.1.1. Halaman Menu Utama

Halaman ini digunakan sebagai tempat untuk menampung semua pilihan-pilihan yang terdapat di dalam sistem yang dirancang seperti terlihat di bawah ini.

Gambar 4.1. Halaman Utama

4.1.2. Halaman Dokumen Dataset

Halaman dokumen dataset merupan halaman data latih, akan dijadikan data pembanding dengan dokumen uji, pada halaman ini terdiri dari fitur tambah yang berfungsi sebagai menambahakan dokumen dan edit untuk melakukan perubahana data dokumen, halaman dokumen dataset dapat dilihat pada gambar 4.2.

Gambar 4.2 Halaman Dokumen Dataset

Pada halaman dataset jika user milih tombol tambah dokumen makan akan membuka halaman form input dokumen seperti gambar 4.3 dan jika user milih edit maka user akan membuka halaman form edit dataset.

Gamabar 4.3. Halaman Form Dokumen Dataset

4.1.3. Halaman Form Uji Dokumen

Pada halaman form uji user harus mengisi dokumen uji dengan input dokumen atau dengan mengimport dokumen file (*.docx) seperti gambar dibawah ini.

Gambar 4.4. Halaman Form Uji Dokumen

4.1.4. Halaman Perangkingan Kesamaan Dokumen Uji

Pada halaman ini user diminta memilih dokumen yang akan dibandingkan dengan dokumen uji. Halaman form pilih dokumen banding dapat dilihat pada gambar 4.5.

Gambar 4.5. Halam Form Dokumen Kesamaan

4.1.5. Halaman Hasil

Menu hasil testing berfungsi untuk melakukan analisis pendeteksian plagiarisme penerapan dari algoritma *Jaro-Winkler distance*.

Selamat Datang

Usage Document:

Dokumen Uji		Dokumen Banding	
No	Term	No	Term
1	pengetahuan	1	pengetahuan
2	tentang	2	tentang
3	tren	3	tren
4	topik	4	topik
5	skripsi	5	skripsi
6	mahasiswa	6	mahasiswa
7	di	7	di
8	sebuah	8	sebuah
9	universitas	9	universitas
10	pada	10	pada
11	umumnya	11	umumnya
12	maupun	12	maupun
13	program	13	program
14	studi	14	studi
15	tertentu	15	tertentu
16	khususnya	16	khususnya
17	dapat	17	dapat
18	membawa	18	membawa
19	manfaat	19	manfaat
20	yang	20	yang
21	sangat	21	sangat
22	positif	22	positif
23	bagi	23	bagi
24	pengembangan	24	pengembangan
25	kurikulum	25	kurikulum
26	perencanaan	26	perencanaan
27	roadmap	27	roadmap
28	penelitian	28	penelitian
29	skala	29	skala
30	institusi	30	institusi
31	namun	31	namun
32	teknologi	32	teknologi
33	untuk	33	untuk
34	secara	34	secara
35	cepat	35	cepat

Gambar 4.6. Halaman Hasil Pendeteksian

5. Kesimpulan dan Saran

5.1.Kesimpulan

Kesimpilan yang dapat di ambil dari penelitian pengecekan kemiripan dokumen dengan algoritma Jaro-Winkler *Disatance* adalah sebagai berikut.

1. Aplikasi ini dapat menerapan algoritma *Jaro-Winkler distance* dalam sebuah sistem pendeteksian terhadap dokumen teks berbahasa Indonesia.
2. Aplikasi ini mampu menampilkan dokumen termirip.

5.2.Saran

1. Agar aplikasi ini dapat diterapkan berbasis android sehingga dapat dilakukan pengembangan pada aplikasi pendeteksi plagiat ini.
2. Memperbarui *user interface* serta menambahkan fitur-fitur yang dapat memperbaiki kinerja aplikasi ini.
3. Agar hasil lebih valid analisis pendeteksian

plagiat ini melibatkan ahli bahasa.

DAFTAR PUSTAKA

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

[2.] Hariyanto, Bambang 2009, Sistem Operasi, Bandung, Informatika.

[3.] Junaedi, Fajar. 2005. Panduan Lengkap Pemrograman PHP untuk Membuat WEB Dinamis. Yogyakarta:PD. Anindya

[4.] Knuth, E. 1973. *The Art of Computer Programming Second Edition Volume I*. Addison-Wesley.

[5.] Nugroho, Adi. 2006. E-commerce. Informatika Bandung. Bandung.

[6.] Sulhan, Mohammad.2007. Pengembangan Aplikasi Berbasis Web dengan PHP & ASP. Yogyakarta: Gava Media.

[7.] Taufikurohman. I, 2015, Pengertian Dokumen dan Jenis-jenis Dokumen, http://irfan-t.heck.in/pengertian-dokumen.xhtml (diakses 20 April 2016).

[8.] Tia Septiana Widi, 2012. Plagiarisme. http://tiaseptianawidi.blogspot.com/2012/02/plagiarisme.html (diakses 19 April 2016).