PENDETEKSIAN TINGKAT SIMILARITAS DOKUMEN BERBASIS WEB MENGGUNAKAN ALGORITMA *WINNOWING*

Nur Fadillah Ulfa[1], Metty Mustikasari[2], Irwan Bastian[3]
[1, 2, 3]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma, Jakarta
[1]nurfadillahulfa@gmail.com,[2]metty@staff.gunadarma.ac.id,
[3]bastian@staff.gunadarma.ac.id

Abstrak

Untuk mengantisipasi plagiarisme dibutuhkan suatu cara yang dapat menganalisis tingkat similaritas kata di antaranya adalah teknik winnowing. Penelitian ini bertujuan untuk membangun sebuah aplikasi berbasis website menggunakan algoritma Winnowing untuk mencari kesamaan pada dua dokumen teks yang diuji. Algoritma yang digunakan untuk mencari nilai hash dalam Winnowing adalah rolling hash. Nilai hash merupakan nilai numerik yang terbentuk dari perhitungan ASCII (American Standard Code for Information Interchange) tiap karakter. Pada hasil eksperimen pengukuran kemiripan ini disimpulkan bahwa semakin kecil tingkat persentase kesamaan dokumen teks yang diuji, maka tingkat kemiripan dokumen kecil dan tidak termasuk plagiat, tetapi jika hasil dari pengujian pada dua dokumen semakin besar, maka dokumen tersebut mempunyai tingkat kemiripan yang tinggi dan tindakan tersebut dianggap plagiat. Penelitian ini juga menambahkan perbandingan nilai k-gram, basis (bilangan prima), nilai window, keterangan persentase, dan kategori plagiarisme.

Kata kunci : *Winnowing, Jaccard, Similaritas, Plagiarisme*

1. Pendahuluan

Pesatnya pekembangan internet menyebabkan semakin banyaknya informasi yang tersedia. Praktek dokumen *plagiarisme* atau penjiplakan selalu menjadi sorotan terutama di kalangan akademis. Tindakan *plagiat* yang dilakukan seseorang ini tidak mencerminkan sikap kreatif sebagai kaum intelektual. *Plagiarisme* dapat dilakukan dengan cara menulis ulang secara lengkap teks yang sudah ada, menghilangkan atau menambahkan kata, parafrase, mengganti kata dengan sinonimnya, mengubah kalimat aktif menjadi pasif atau sebaliknya (Hutami, 2012). Dengan melakukan penyalinan dan modifikasi dengan sempurna, maka dapat menyamarkan *plagiarisme* karena karya yang dihasilkan hampir seperti karya milik sendiri sehingga sulit untuk mengetahui keoriginalitasan dari sebuah dokumen. Kemungkinan untuk mendeteksi *plagiarisme* terhadap sekumpulan dokumen yang sangat banyak akan membutuhkan waktu yang sangat lama dan akan mengalami kesulitan. Suatu karya ilmiah atau penulisan seseorang dikatakan sebagai hasil penjiplakan apabila kutipan yang dilakukan tidak disertai penyebutan referensi secara benar. Oleh karena itu dibutuhkan sebuah sistem yang dapat mendeteksi plagiat (Elbegbayan, 2005).

Pada penelitian ini akan dibangun pendeteksian tingkat similaritas dokumen berbasis web menggunakan Algoritma *Winnowing*. Untuk membuat web ini digunakan aplikasi *Xampp* dan PHP. Sistem ini dapat melakukan pendeteksian kemiripan (*similarity*) dengan membandingkan dokumen asli dengan dokumen uji yang keduanya merupakan dokumen berbahasa Indonesia menggunakan algoritma *Winnowing*. *Winnowing* merupakan algoritma *document fingerprinting* yang

akurat dalam mengidentifikasi penyalinan teks termasuk bagian kecil yang mirip dalam sekumpulan dokumen melalui *fingerprint*yang dihasilkan (Schleimer, Wilkerson & Aiken , 2003).

Algoritma *Winnowing*, diawali dengan pembagian *k-gram*, penghitungan *hashing*, *windowing*, kemudian proses mendapatkan *fingerprint*. Setelah mendapatkan *fingerprint*kemudian dilakukan penghitungan persentase *similarity* antar dokumen (Pratama, Cahyono, & Marthasari, 2011).

Aplikasi pendeteksian tingkat kemiripan dokumen teks ini dilengkapi dengan pengukuran terhadap persentase kemiripan dokumen teks dengan Algoritma *Winnowing* berdasarkan parameter nilai *k-gram*, nilai basis *hash*, dan nilai ukuran *window*.

Adapun batasan masalah dari penelitian ini yaitu:

1. Data yang diuji bertipe teks dan .doc .
2. Nilai *k-gram*, nilai basis *hash* dan nilai ukuran *window* saat pengujian ditentukan oleh user.
3. Pada proses *rolling hash*, menggunakan nilai ASCII (*American Standard Code for Information Interchange*) dari karakter.
4. Maksimal kata dalam dokumen uji 4500 kata.
5. Maksimal dokumen yang bisa di upload adalah 10,9 MB.
6. Nilai *k-gram* dan nilai ukuran *window* yang digunakan dibatasi antara 1 sampai 12.
7. Nilai basis *hash* yang digunakan dibatasi antara bilangan prima 2 sampai 30.

Tujuan yang ingin dicapai dalam penelitian ini adalah membuat sebuah aplikasi yang dapat mengukur tingkat kemiripan dokumen teks menggunakan Algoritma *Winnowing* berdasarkan parameter nilai *gram*, nilai basis *hash* dan nilai ukuran *window.*

2. Tinjauan Pustaka

2.1 Algoritma *Winnowing*

Winnowing adalah algoritmayang digunakan untuk melakukan proses pengecekkan kesamaan kata(*document fingerprinting*) untuk mengidentifikasi *plagiarisme*(Mozgovoy, 2007). Untuk melakukan pendeteksian dokumen dengan algoritma *winnowing* diperlukan beberapa rumus, antara lain :

1. Mencari nilai *hash* pertama.

$$H_{(C1\ldots CK)} = c1 * b^{(k-1)} + c2 * b^{(k-2)} + \cdots + c^{(k-1)} * b + ck \quad (1)$$

2. Mencari nilai *hash* kedua.

$$H_{(c2\ldots ck+1)} = (H_{(c1\ldots ck)} - c1 * b^{(k-1)}) * b + c^{(k+1)} \quad (2)$$

3. Pengukuran nilai similaritas dengan menggunakan *Jaccard Coefficient.*

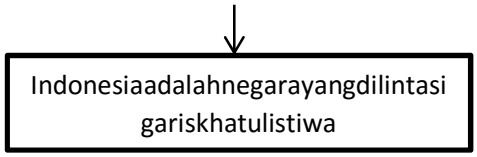
$$\text{Similaritas } (d_i\ d_j) = \frac{|w(d_i) \cap (d_j)|}{|w(d_i) \cup (d_j)|} \quad (3)$$

Dalam melakukan pendeteksian penjiplakan terdapat kebutuhan mendasar yang harus dipenuhi oleh suatu algoritma penjiplakan yaitu :*Whitespace Insensitivity, Noise Surpression,* dan *Position Independence.* Arti dari *Whitespace Insensitivity* adalah melakukan pencocokan terhadap file teks seharusnya tidak terpengaruh oleh spasi, jenis huruf (kapital atau normal), tanda baca dan sebagainya. *Noise Surpression* adalah menghindari penemuan kecocokan dengan panjang kata yang terlalu kecil atau kurang relevan, misal: ‘the’. Panjang kata yang ditengarai merupakan penjiplakan harus cukup untuk membuktikan bahwa kata-kata tersebut telah dijiplak dan bukan merupakan kata yang umum digunakan. Serta *Position Independence*, yang berarti penemuan kecocokan / kesamaan tidak harus bergantung pada posisi kata-kata. Walau tidak dalam berada posisi yang sama pencocokan juga harus dilakukan.

Pada detil bahasan berikut ditunjukkan keterkaitan antara teori-teori diatas dengan langkah-langkah yang dilakukan dalam algoritma *Winnowing*. Proses untuk menghasilkan *fingerprint*sebuah dokumen adalah sebagai berikut :

1. Membuang karakter-karakter tidak relevan seperti tanda baca.
   Contoh teks : Indonesia adalah Negara yang dilintasi garis khatulistiwa.

Indonesia adalah Negara yang dilintasi garis khatulistiwa



2. Membentuk rangkaian *k-gram* dari teks, semisal k=4. Pemotongan *k-gram* berdasarkan nilai k.
3. Membentuk rangkaian *k-gram* dari teks, semisal k=4. Pemotongan *k-gram* berdasarkan nilai k.

Dengan k=4, maka *k-gram* yang dihasilkan sebesar: 48 gram.

4. Melakukan fungsi *hash* untuk setiap *k-gram*.

indo ndon done ones nesi esia siaa iaad aada adal dala alah lahn ahne hneg nega egar gara aray raya ayan yang angd ngdi gdil dili ilin lint inta ntas tasi asig siga igar gari aris risk iskh skha khat hatu atul tuli ulis list isti stiw tiwa

154276 159841 147842 162277 160001 149598 166934 152659 142041 142382 146122 143346 156739 143002 152948 159861 148075 150181 144089 164899 144925 174101 143650 160078 150456 147098 154088 157779 154438 161628 167503 144280 167000 153399 150189 144171 165811 154951 167253 156184 151554 144538 169846 170065 157834 155051 168375 168507

Persamaan (1) adalah perhitungan fungsi hash dari algoritma *Winnowing* dengan c sebagai nilai ASCII, b sebagai nilai basis bilangan prima, dan banyak karakter k (Winangga, 2014):

$$H_{(C1\ldots CK)} = c1 * b^{(k-1)} + c2 * b^{(k-2)} + \cdots + c^{(k-1)} * b + ck \quad (1)$$

Untuk menghitung *hash* dari *k-gram* *c*2…*c*k+1, dapat menggunakan persamaan (2) :

$$H_{(c2\ldots ck+1)} = (H_{(c1\ldots ck)} - c1 * b^{(k-1)}) * b + c^{(k+1)} \quad (2)$$

Sebagai contoh *k-gram* dari "*indo*" dengan nilai *b* = 11 dan *k* = 4memiliki nilai *hash*:

$$H_{(indo)} = ascii_{(i)} * 11^{(3)} + ascii_{(n)} * 11^{(2)} + ascii_{(d)} * 11^{(1)} + ascii_{(o)} * 11^{(0)}$$

$$H_{(indo)} = (105*1331) + (110*121) + (100*11) + (111*1) = 154276$$

5. Membentuk nilai ukuran *window* dari nilai-nilai *hash* dengan ukuran 4. Berikut ini adalah hasil dari *window*:

154276 159841 **147842** 162277

160001**149598** 166934 152659

142041 142382 146122 143346

156739 **143002** 152948 159861

148075 150181 **144089** 164899

144925 174101**143650** 160078

150456**147098** 154088 157779

154438 161628 167503**144280**

167000 153399 150189 **144171**

165811 154951 167253 156184

151554 **144538** 169846 170065

157834 155051 168375 168507

Kemudian *fingerprint*yang dihasilkan adalah sejumlah 15 nilai *hash* dari 15 *window* yaitu: 14784214959814204114238214612214334614300214807514408914492514365014709 8144280144171 144538.

6. Langkah kelima yaitu memilih nilai *hash* terkecil dari setiap *widow* untuk dijadikan sebagai *fingerprint.*

   Hasil *fingerprint* Teks 1 :

   [147842,2] [149598,5]

   [142041,8] [142382,9] [143002,13] [148075,16]

   [144089,18] [143650,22]

   [147098,25] [154088,26]

   [154438,28] [144280,31] [144171,35] [154951,37] [151554,40] [144538,41] [155051,45]

   Teks 2 : indonesia adalah negara yang dilintasi garis khatulistiwa atau ekuator.

   Hasil *fingerprint* Teks 2 :

   [[147842,2] [149598,5]

   [142041,8] [142382,9]

   [143002,13] [148075,16][144089,18] [143650,22][147098,25] 154088,26]

   [154438,28] [144280,31] [144171,35] [154951,37] [151554,40] [144538,41] [155051,45] [155318,48] [142217,50] [144327,51]

   [144482,53] [148762,55] [144478,58]

7. Langkah keenam yaitu pengukuran nilai similaritas dengan menggunakan *Jaccard Coefficient.*

$$\text{Similaritas } (d_i\, d_j) = \frac{|w(d_i) \cap (d_j)|}{|w(d_i) \cup (d_j)|} \quad (3)$$

S = x 100% = 73.9%

2.2 Literature Review

Penelitian yang dilakukan oleh Elegbayan (2005) membuat studi literatur tentang *Winnowing*, sebuah algoritma *fingerprint* untuk dokumen. Algoritma *winnowing* ini memilih *fingerprint* dari nilai hash, nilai *k - gram*, yang berdampingan dengan *substring* panjang k. Penelitian ini juga menunjukkan dokumen *fingerprint* dan contoh untuk menunjukkan kinerja algoritma. Dalam penelitian ini telah membahas tentang *fingerprint*, sebuah teknik perlindungan hak cipta dan contoh metodenya dengan keunggulan dan kelemahannya. Norzima mengambil dokumen *fingerprinting* sebagai sebuah kasus dan telah menunjukkan bahwa meskipun dokumen *plagiarisme* adalah yang paling sulit untuk dideteksi, ada batas tertentu ketika menggunakan hash dari *k-gram* untuk memilih dokumen *fingerprint*. Norzima juga telah menyajikan *Winnowing*, algoritma *fingerprinting* dokumen yang efisien dan cocok untuk pendeteksian similaritas dokumen.

Penelitian yang dilakukan oleh Schleimer, Wilkerson, & Aiken (2003) memperkenalkan kelas algoritma dokumen *fingerprinting* lokal, yang tampaknya untuk mengambil properti penting dari *fingerprint*. Setiap teknik terjamin untuk mendeteksi dokumen salinan. Saul membuktikan dokumen yang diuji pada kinerja algoritma lokal. Saul juga mengembangkan *Winnowing*, algoritma *fingerprint* lokal efisien, dan menunjukkan bahwa kinerja *winnowing* dalam 33% yang lebih rendah. Akhirnya, kami juga memberikan hasil percobaan pada data

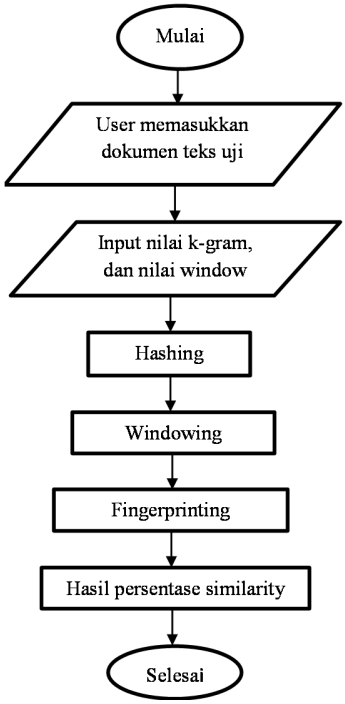
w*eb*, Pengalaman laporan dengan MOSS, dan memberikan layanan untuk pendeteksian dokumen *plagiarisme*. Penelitian ini telah mempresentasikan hal yang tidak biasa untuk batas bawah pada kompleksitas dari setiap dokumen *fingerprint* lokal algoritma. Akhirnya, kita telah membahas serangkaian eksperimen yang menunjukkan efektivitas *winnowing* data yang nyata, dan kami memiliki laporan pengalaman menggunakan *winnowing* (*Synonym Recognition*). Penelitian ini membahas tentang sistem deteksi duplikasi menggunakan algoritma *winnowing* yang outputnya berupa seperangkat nilai *hash*sebagai *fingerprint*ing dokumen yang diperoleh melalui metode *k-gram*. Masukan dari proses dokumen *fingerprinting* adalah file teks. Maka output-nya akan menjadi set nilai *hash*, yang disebut *fingerprint*.

Penelitian yang dilakukan oleh Pratama, Cahyono, &Marthasari (2011) membuat perbandingan antara file teks yang telah dimasukkan. Adanya pengakuan konsep sinonim dimaksudkan untuk dapat mengenali kata-kata yang mengandung sinonim sebagai tindakan *plagiarisme*. Mendeteksi duplikat menggunakan sinonim mendapatkan persentase lebih tinggi daripada tanpa menggunakan sinonim. Kesimpulannya adalah tindak penjiplakan dapat dilakukan dengan *modify* yang mana dengan mengubah beberapa bagian bahkan keseluruhan, yaitu dengan mengubah kata-kata dengan sinonim. Mendeteksi duplikasi menggunakan sinonim mendapatkan hasil persentase yang lebih tinggi daripada tanpa menggunakan sinonim.

3. Analisis dan Pembahasan

3.1 Rancangan Sistem yang diusulkan

Pada awal antar muka, sistem ini akan memberikan pilihan kepada user untuk memasukkan teks yang akan diuji kesamaannya. Setelah itu sistem akan menganalisis persentase kemiripan.



Gambar 1.Flowchart Alur Sistem

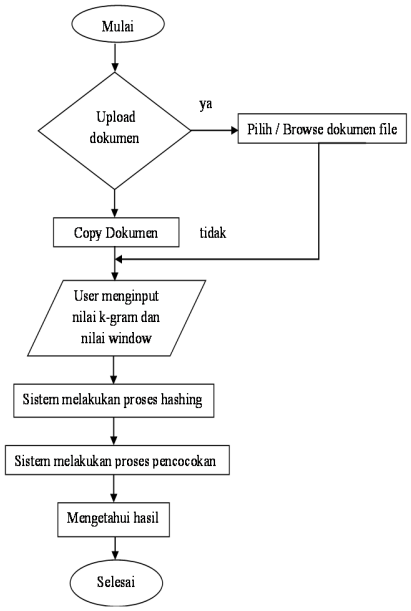
Data yang diuji dalam sistem ini adalah berupa teks. Dokumen teks dapat bertipe .txt. *Flowchart* sistem dapat dilihat pada Gambar 1.Dimulai dari user memasukkan dokumen teks uji, tahap kedua adalah user menginput nilai *k-gram* dan nilai *window*, tahap ketiga yaitu aplikasi mulai melakukan proses *hashing*, tahap keempat aplikasi melakukan proses *windowing*, tahap kelima aplikasi juga melakukan proses *fingerprinting*, lalu tahap keenam user akan mengetahui hasil *similarity* dari kedua dokumen yang telah diuji, kemudian tahap ketujuh user juga mengetahui hasil dari persentase *similarity* dua dokumen tersebut, kemudian tahap terakhir adalah *stop* untuk mengakhiri proses *flowchart*.

Data yang dibutuhkan dalam penelitian ini data berbentuk teks tanpa gambar. Data dokumen penelitian diambil dari internet. Dokumen latih yang digunakan merupakan hasil pengubahan dari dokumen uji dengan beberapa kondisi sebagai berikut:

1. Kategori Nihil (0%) yaitu kedua dokumen tidak terindikasi plagiat karena benar-benar berbeda baik dari segi isi dan kalimat secara keseluruhan.
2. Kategori Sedikit Kesamaan (<15%) yaitu kedua dokumen hanya mempunyai sedikit kesamaan.
3. Kategori Plagiat Sedang (15-50%) yaitu kedua dokumen terindikasi plagiat tingkat sedang.
4. Kategori Mendekati Plagiarisme (>50%) yaitu kedua Hasil uji menunjukkan lebih dari 50%, dapat dikatakan bahwa dokumen yang diuji mendekati tingkat plagiarisme.
5. Kategori Plagiarisme (100%) yaitu kedua Dokumen uji dapat dipastikan murni plagiat karena dari awal dan sampai akhir isi dokumen adalah sama.

3.2 Rancangan *User Interface*

Pada rancangan *user interface*, sistem ini akan memberikan penjelasan tentang alur kerja *user* dan sistem. *Flowchart* sistem dapat dilihat pada Gambar 2.



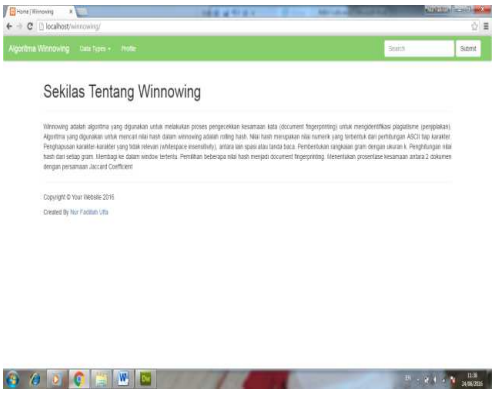
Gambar 2.Flowchart User Interface

Pada *flowchart user interface*, pertama *user* akan diberi pilihan untuk mengupload dokumen atau *copy* dokumen. Jika upload dokumen maka akan masuk pilihan atau *browse* dokumen file. Jika *copy* dokumen maka akan langsung ke tahap input nilai *k-gram* dan nilai *window*.

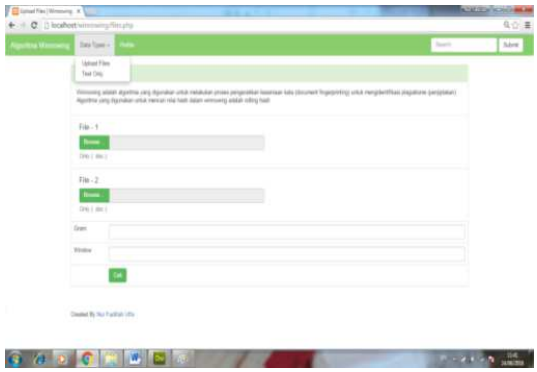
Kemudian sistem akan melakukan proses *hashing*. Lalu sistem akan melakukan proses pencocokan. Setelah itu user dapat mengetahui hasil, kemudian *user* bisa keluar dari aplikasi.

3.3 Implementasi

Tampilan aplikasi merupakan interaksi antara sistem dengan *user*. Tampilan aplikasi tidak hanya harus menarik, namun juga harus mempermudah user dalam menggunakan aplikasi tersebut. Tahap awal dalam merancang aplikasi ini yaitu membuat tampilan awal pada *website*. Tampilan awal terdiri dari sekilas tentang pengertian Algoritma *Winnowing* yang dibuat menggunakan *div tag*. Untuk lebih jelas dapat dilihat pada Gambar 3



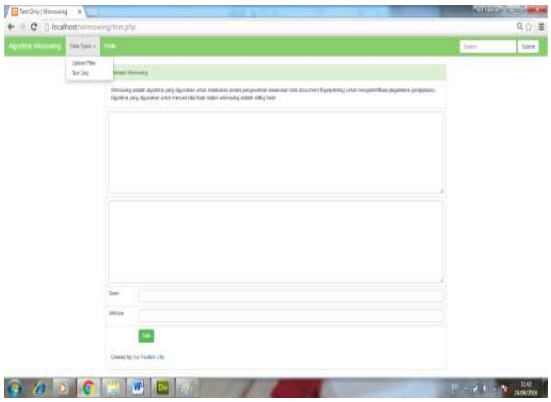
Gambar 3. Tampilan Halaman Awal



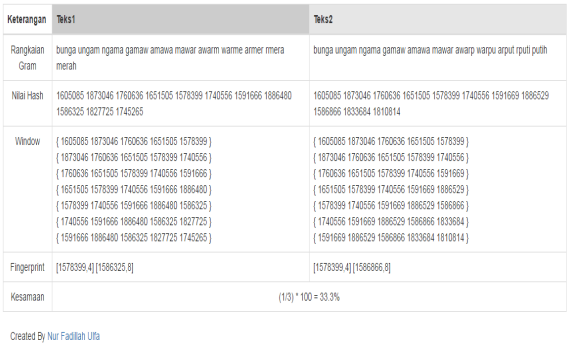
Gambar 4. Halaman Upload files

Setelah pembuatan tampilan awal *website*, selanjutnya membuat tampilan untuk *upload* dokumen dan *text only*. Pada tampilan kedua ini terdapat penjelasan tentang *winnowing* yang dibuat menggunakan *div class penel-heading* dan *div class panel-body*, kemudian tulisan file-1 dan file-2 dibuat menggunakan *div class*. Lalu tombol *browse* dibuat menggunakan *button* dan di samping tombol *browse* dibuat menggunakan label. Selanjutnya *gram* dan *window* dibuat menggunakan *div class*. Lalu tombol Cek dibuat meggunakan *button*. Untuk lebih jelas dapat dilihat pada Gambar 4.

Selain bisa menguji dokumen dengan cara *Upload files, website* ini juga bisa menguji dokumen dengan cara *Text only* atau *copy* ke dalam kolom *website* yang sudah tersedia. Pada tampilan di *Text only* ini terdapat penjelasan tentang *winnowing* yang dibuat menggunakan *div classpenel-heading* dan *div class panel-body*, kemudian kolom kosong yang akan di isi teks dibuat menggunakan *Text area*.Selanjutnya *gram* dan *window* dibuat menggunakan *div class*. Kolom disamping *gram* dan *window* dibuat menggunakan label. Lalu tombol Cek dibuat meggunakan button. Untuk lebih jelas dapat dilihat pada Gambar 5.



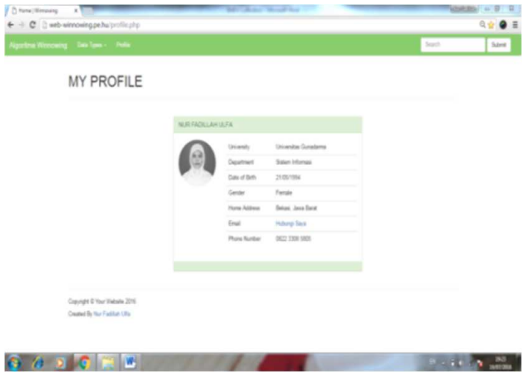
*Gambar 5.*Halaman Text only



Keterangan	Teks1	Teks2
Rangkaian Gram	bunga ungam ngama gamaw amawa mawar awarm warme armer rmera merah	bunga ungam ngama gamaw amawa mawar awarp warpu arput rputi putih
Nilai Hash	1605085 1873046 1760636 1651505 1578399 1740556 1591666 1886480 1586325 1827725 1745265	1605085 1873046 1760636 1651505 1578399 1740556 1591669 1886529 1586866 1833684 1810814
Window	{ 1605085 1873046 1760636 1651505 1578399 } { 1873046 1760636 1651505 1578399 1740556 } { 1760636 1651505 1578399 1740556 1591666 } { 1651505 1578399 1740556 1591666 1886480 } { 1578399 1740556 1591666 1886480 1586325 } { 1740556 1591666 1886480 1586325 1827725 } { 1591666 1886480 1586325 1827725 1745265 }	{ 1605085 1873046 1760636 1651505 1578399 } { 1873046 1760636 1651505 1578399 1740556 } { 1760636 1651505 1578399 1740556 1591669 } { 1651505 1578399 1740556 1591669 1886529 } { 1578399 1740556 1591669 1886529 1586866 } { 1740556 1591669 1886529 1586866 1833684 } { 1591669 1886529 1586866 1833684 1810814 }
Fingerprint	[1578399.4] [1586325.8]	[1578399.4] [1586866.8]
Kesamaan	(1/3) * 100 = 33.3%	

Created By Nur Fadilah Ulfa

*Gambar 6.*Tampilan Halaman Keterangan



Gambar 7. Halaman My Profile

Tampilan selanjutnya adalah tampilan keterangan teks 1 dan teks 2 yang sudah di uji. Tampilan keterangan ini dibuat menggunakan tabel yaitu tabel *table-bordered*. Untuk lebih jelas dapat dilihat pada Gambar 6

Tampilan selanjutnya adalah tampilan *My Profile*. tampilan halaman *My profile* ini terdapat penjelasan tentang data diri pembuat website. Terdiri dari tabel dan kolom-kolom berisi nama, alamat, universitas, tanggal lahir, jenis kelamin, email, dan nomor telephone. Untuk lebih jelas dapat dilihat pada Gambar 7.

Tabel 1. Hasil Uji Dokumen

Dokumen		k-gram	Basis	Window	Keterangan	Kategori
Dok 1 (1,41 MB)	Dok 2 (1,41 MB)	4	3	4	100%	Plagiarisme atau isi dokumen sama persis
Dok 1 (3,19 MB)	Dok 2 (3,19 MB)	3	2	3	>50%	Mendekati plagiarisme atau isi dokumen hampir sama
Dok 1 (4,20 MB)	Dok 2 (4,20 MB)	3	2	3	15-50%	Plagiat sedang, atau isi dokumen 50% sama
Dok 1 (6,30 MB)	Dok 2 (6,30 MB)	4	3	4	<15%	Isi dokumen Sedikit kesamaan
Dok 1 (10MB)	Dok 2 (10 MB)	5	4	5	0%	Tidak ada kesamaan atau isi dokumen berbeda

Pada Tabel 3 telah dilakukan perbandingan pengecekan similaritas kata pada dokumen. Dalam percobaan ini digunakan dokumen 1 dan dokumen 2. Kategori *plagiarisme* yaitu dokumen uji dapat dipastikan murni plagiat karena dari awal dan sampai akhir isi dokumen adalah sama, dilakukan percobaan dan input nilai *k-gram*=4, nilai basis (bilangan prima)=3, dan nilai *window*=4. Kategori mendekati *plagiarisme* yaitu kedua Hasil uji menunjukkan lebih dari 50%, dapat dikatakan bahwa dokumen yang diuji mendekati tingkat *plagiarisme* isi dokumen tersebut juga hampir sama. input nilai *k-gram*=3, nilai basis (bilangan prima)=2, dan nilai *window*=3.

Kategori plagiat sedang yaitu kedua dokumen terindikasi plagiat tingkat sedang isi dokumen juga 50% kesamaannya. dilakukan percobaan input nilai *k-gram*=3, nilai basis (bilangan prima)=2, dan nilai *window*=3. Kategori sedikit kesamaan yaitu kedua dokumen hanya mempunyai sedikit kesamaan. dilakukan percobaan input nilai *k-gram*=4, nilai basis (bilangan prima)=3, dan nilai *window*=4. Kategori tidak ada kesamaan yaitu kedua dokumen tidak terindikasi plagiat karena benar-benar berbeda baik dari segi isi dan kalimat secara keseluruhan. dilakukan percobaan input nilai k-gram=5, nilai basis (bilangan prima)=4, dan nilai window=5.

3.4 Perbandingan eksekusi dokumen dan waktu

Dalam menganalisis dokumen dibutuhkan uji coba perbandingan ukuran dokumen untuk mengetahui berapa lama waktu yang dibutuhkan aplikasi berbasis *website* ini menjalankan eksekusi *time*. Berikut ini adalah tabel hasil uji coba perbandingan ukuran dokumen dan berapa lama waktu yang digunakan.

Tabel 2. Perbandingan Eksekusi Dokumen Dan Waktu

Ukuran Dok 1	Ukuran Dok 2	K-gram	Window	Waktu
1,18 MB	1,12 MB	4	4	14 detik
2,48 MB	2,30 MB	5	5	19 detik
3,51 MB	3,08 MB	6	6	25 detik
4,61 MB	4,31 MB	7	7	42 detik
5,90 MB	5,74 MB	8	8	2 menit 10 detik
6,35 MB	6,35 MB	3	3	10 menit 48 detik
7,11 MB	7,01 MB	4	4	12 menit

				11 detik
8,45 MB	8,42 MB	5	5	14 menit 38 detik
9,06 MB	9,04 MB	4	4	16 menit 42 detik
10,9 MB	10,7 MB	3	3	18 menit 58 detik

4. Penutup

Aplikasi pendeteksian tingkat kemiripan dokumen berbasis *web* menggunakan algoritma *winnowing* telah berhasil dibangun.*Website* Algoritma *Winnowing* juga telah diunggah ke idHostinger dengan alamat www.web-*winnowing*.pe.hu pada tanggal 09 Juli 2016. Aplikasi website ini memberikan hasil berupa persentase dan keterangan bahwa kedua dokumen yang diuji termasuk plagiat atau tidak.Waktu proses untuk pendeteksian ini lebih lama jika memproses file yang cukup besar. Semakin besar nilai k-gram maka hasilnya tidak akurat. Pada hasil eksperimen pengukuran kemiripan ini disimpulkan bahwa semakin besar persentase kesamaan dokumen teks yang diuji, maka tingkat kemiripan dokumen tinggi dan tindakan tersebut dianggap plagiat. Untuk penelitian lebih lanjut diharapkan agar dokumen yang berekstensi .pdf juga bisa diproses untuk kemudahan pengguna.

Daftar Pustaka

Elbegbayan, N (2005). *Winnowing, a Document Fingerprinting Algorithm*, Department of Computer Science, Linkoping University.

Kurniawati, A, Wicaksana, I W. Simri (2008). Perbandingan Pendekatan Deteksi Plagiarism Dokumen Dalam Bahasa Inggris.*KOMMIT 2008*. Dipresentasikan pada : Seminar Ilmiah Nasional Komputer dan Sistem Intelijen, Jakarta. Jakarta :Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma.

Mozgovoy, M. (2007). *Enhancing Computer-Aided Plagiarism Detection. (*Dissertation. Computer Science and Statistics. University of Joensuu 2007). University of Joensuu. 2007.

Hutami, R.R., Suyanto (2012). *Implementasi sistem pendeteksian plagiat pada dokumen bahasa Indonesia menggunakan Algoritma Winnowing*. (Skripsi S1, Jurusan Teknik Informatika Universitas Telkom 2012). Dari Open Library Universitas Telkom 2012.

Pratama, M.R, Cahyono, E.B, Indah Marthasari, Gita, (2011). *Aplikasi Pendeteksi Duplikasi Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Winnowing Dengan Metode K-Gram Dan Synonym Recognitio.* (Skripsi S1 Jurusan Teknik Informatika Universitas Muhammadiyah Malang). Dari UMM Institutional Repository

Schleimer, S.. Wilkerson,D.S, & Aiken,A. (2003). Winnowing: local algorithms for document fingerprinting. *Proceedings of the 2003 ACM SIGMOD international conference on Management of data, 76-85*. Dipresentasikan pada SIGMOD international conference on Management of data. New York, NY, USA: ACM. DOI=http://dx.doi.org/10.1145/872757.872770

Winangga, M. (2014). Deteksi Plagiarisme pada Dokumen Teks Bahasa Indonesia menggunakan Algoritma Winnowing dengan Stemming. *Repositori Jurnal Mahasiswa PTIIK UB.* Jurusan Teknik Informatika Universitas Brawijaya. Diakses online: http://filkom.ub.ac.id/doro/archives/detail/DR00123201406

BiodataPenulis

Nur Fadillah Ulfa, memperoleh gelar S1 Jurusan Sistem Informasi di Universitas Gunadarma.

Metty Mustikasari, memperoleh gelar S1 Jurusan Sistem Informasi di Universitas Gunadarma. Memperoleh gelar S2 Computer Science di Curtin University Australia. memperoleh gelar S3 Jurusan Teknologi Informasi di Universitas Gunadarma. Saat ini menjadi pengajar di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Gunadarma.

Irwan Bastian, memperoleh gelar S1 Jurusan Sistem Informasi di Universitas Gunadarma. Memperoleh gelar S2 MMSI Magister Pascasarjana di Universitas Gunadarma. Saat ini sedang menyelesaikan S3 Jurusan Teknologi Informasi di Universitas Gunadarma. Menjadi pengajar di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Gunadarma.

BERITA ACARA PELAKSANAAN HASIL SEMINAR SESI PARALEL KNASTIK 2016

Judul : Pendeteksian Tingkat Similaritas Dokumen Berbasis Web Menggunakan Algoritma Winnowing

Pemakalah : Nur Fadillah Ulfa, Metty Mustikasari, Irwan Bastian

Moderator : Gloria Virginia, S.Kom., MAI, Ph.D

Notulis : Yoas

Peserta : 11 orang di ruang : D.3.2

Tanya Jawab :

Pertanyaan (oleh sdr. Hendri):

1. Saya melihat penggunaan *K-Gram* bervariasi. Bukankah lebih baik menggunakan 1 dokumen dan dieksekusi dengan 1 *gram*, setelah ditemukan *gram* yg terbaik bisa dieksekusi dengan dokumen yang lain, Karena kalua menggunakan nilai yang demikian similaritas menjadi rancu. Kenapa tidak memakai nilai gram yang *fix*?
2. Setiap dokumen pasti memiliki *noise*, namun disitu saya tidak melihat bagaimana metode membersihkan *noise* pada dokumen?

Jawaban:

1. *K-Gram* masih diinput sesuai keinginan pengguna, penyaji belum meneliti lebih lanjut.
2. *Noise* yang dibersihkan adalah berupa tanda baca, spasi, dan huruf yang disetarakan menjadi *lowercase*.

Masukan (oleh sdr. Alz):

1. Belum spesifik berapa prosentase *similarity* yang bisa ditetapkan sebagai plagiasi.
2. Penelitian menggunakan metode *Hashing*. *Hashing* memiliki kelemahan dapat menghasilkan dua kunci berbeda pada tulisan atau dokumen yang sama. Hal ini belum dicantumkan pada batasan masalah penelitian.
3. Pengujian akan lebih baik bukan berdasarkan besarnya dokumen/file, melainkan lebih ke isi dokumen.

Masukan Seminar :

Metodologi penelitian masih lemah, perlu ditingkatkan, paling tidak berdasarkan masukkan para penanya.

Yogyakarta, 19 November 2016

Moderator Kelas

Gloria Virginia, S.Kom., MAI, Ph.D.

Penyaji Makalah

Nur Fadillah Ulfa