PEMBUATAN APLIKASI PENGUKURAN TINGKAT KEMIRIPANDOKUMEN BERBASIS WEB MENGGUNAKAN ALGORITMA WINNOWING

***Nur Fadillah Ulfa*[1]**
***Metty Mustikasari*[2]**

[1,2]FakultasIlmuKomputerdanTeknologiInformasiUniversitasGunadarma,
[1]nurfadillahulfa@gmail.com
[2]metty@staff.gunadarma.ac.id

Abstrak

Plagiarisme atau penjiplakan adalah suatu tindakan pengambilan karangan atau pendapat orang lain dan menjadikannya seolah karangan tersebut sebagai pendapat sendiri. Untuk mengantisipasinya, dibutuhkan suatu cara yang dapat menganalisis teknik plagiat yang dilakukan. Algoritma Winnowing merupakan salah satu algoritma pada metode document fingerprinting. Metode ini akurat dalam mengidentifikasi penyalinan teks termasuk bagian kecil yang mirip dalam sekumpulan dokumen melalui fingerprint yang dihasilkan. Melalui pencocokan fingerprint akan diperoleh nilai similarity antar dokumen. Semakin kecil tingkat persentase kesamaan dokumen teks yang diuji, maka dokumen tersebut tidak termasuk plagiat, tetapi jika hasil dari pengujian pada dua dokumen semakin besar, makadisimpulakan bahwa dokumen tersebut menyerupai tindakan plagiat.

Kata kunci :Algoritma *Winnowing*, Dokumen, Similaritas, *Plagiarism.*

DESIGNING WEB-BASED APPLICATION TO MEASURE THE LEVELS OF SIMILARITY OF THE DOCUMENTS USING WINNOWING ALGORITHM

Abstract

Plagiarism is an act of taking essay or opinion of others and make it as the wreath is as his own opinion. In anticipation of this, we need a way to analyze the technique of plagiarism committed. Winnowing algorithm is one of the methods of document fingerprinting algorithms. The result shows that this method is accurate to identify the plagiarism of the text, even small parts that are similar among documents through the generated fingerprint. Through fingerprint, this would obtain the value of similarity between documents. The smaller the similarity percentage level in the document, the less the value of plagiarism found in the document. However, if the results of the tests showed bigger similarity, then it could be concluded that the documents are resembles. This is indicating an act of plagiarism

Keywords : Winnowing Algorithms, Documents, Similarity, Plagiarism

PENDAHULUAN

Pemanfaatanteknologi digital telahmenjadikebutuhandalam era modern saatini.Komponen yang ada di dalamdunia digital salahsatunyaadalahdokumenteks. Dokumendalambentuk digital memudahkandalamhalpenyimpanan, efisien, mudahdicari, bahkanmudahdalamhalpenjiplakan.

Penjiplakanatauplagiarismeberartimencontohataumeniruataumencuritulisandankarya orang lain yang kemudiandiakuisebagaikarangannyasendiridenganataupuntanpaseizinpenulisnya. Plagiarismeberasaldaribahasalatin*plagiarius* yang berartimerampok,ataumembajak. Plagiarismemerupakantindakan pencurianataukebohonganintelektual.Plagiarismeadalahtindakanpenyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, ataumenyatakansebagaimiliksendirisebuahpikiran, ide, tulisan, atauciptaan yang sebe-narnyamilik orang lain [1]. Pen-jiplakandokumen digital bukanlahhal yang susah, cukupdenganmenggunakanteknik*copy-paste-modify* padasebagianisidokumendanbahkankeseluruhanisidokumensudahbisadikatakanbahwadokumentersebutmerupakanhasilduplikasidari do-kumen lain.

Praktekpenjiplakandokumenini seringkaliditerapkanoleh akademisibaiktingkatsekolahmaupunperguruantinggi.Tindakanplagiat yang dilakukanolehsiswamaupunmahasiswainisangattidak mencerminkansikapkreatifdanterpelajarsebagaikaumintelektual. Demi menyelesaikantugas-tugasnyadengancepat, seseorangdapatmelakukanteknik*copy-paste-modify* tanpaperlumempelajaridanmengeksplorasimateriterlebihdahulu. Kadangkalatindakpenjiplakaninidimodifikasidenganmengganti kata-kata yang mengandungsinonim, denganmaksud agar terlihatberbedadaripekerjaanteman. Hal semacaminidapatmenimbulkanmasalahterhadapevaluasihasilbelajarsiswa/mahasiswa.

Kesamaandokumenbukanhanya ditinjaudarisisi kata yang digunakansebagaipenyusunkalimatadalahsama, akantetapi juga dikatakanmiripapabilaisidokumenmemilikimakna yang sama. Penelitianpengukurankesamaandokumen Bahasa Indonesia yang adahanya me-ngukurkesamaan kata ataupun kali-mat, belummempertimbangkanstruk-tur kalimat,jumlahkalimat, posisikalimatdanmakna kata untukmembandingkankalimat [2].

Proses pendeteksian penjiplakan ini menggunakan algoritma *Winnowing* yang mana *output*-nya berupa sekumpulan nilai *hash* yang didapatkan melalui metode *k-gram*. Sedangkan konsep *synonym recognition* ini dimaksudkan untuk dapat mengenali kata-kata yang mengandung sinonim sebagai tindak penjiplakan.

METODE PENELITIAN

AlgoritmaWinnowing

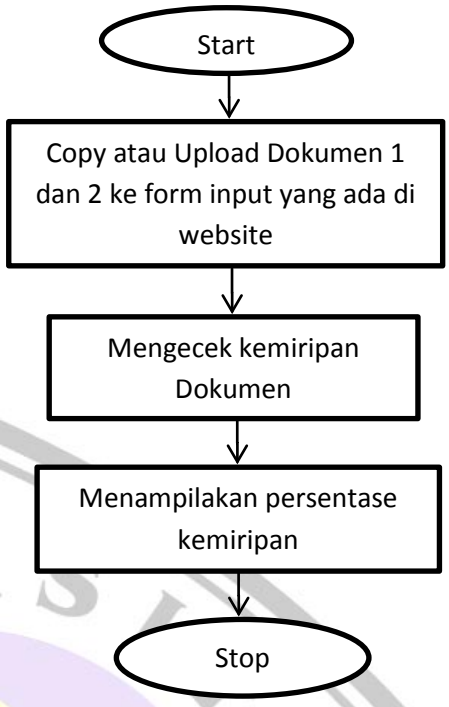
Banyakcaraataumetode yang dapatdigunakanuntukmendeteksipenjiplakandalam file text. Namunada3 kebutuhanmendasar yang harusdipenuhiolehalgoritmadeteksipenjiplakan[3]. Kebutuhanpertama*adalahWhitespace Insensitivity* yang berartidalammelakukanpencocokanterhadap file teksseharusnyatidakterpengaruholehspasi, jenishuruf, tandabacadansebagainya. Kedua, *noise Surpression*yang berartimenghindaripenemuankecocokandenganpanjang kata yang terlalukecilataukurangrelevan, misal: „the". Panjang kata yang ditengaraimerupakan penjiplakanharuscukupuntuk membuktikanbahwa kata-kata

tersebuttelahdijiplakdanbukanmerupakan kata yang umumdigunakan.Ketiga, *position Independence* yang berartipenemuankecocokan/kesamaanharustidakbergantungpadaposisi kata-kata.Meskipunposisinyatidaksama, kecocokanharusdapatditemukan.

Winnowing adalahalgoritma yang digunakanuntukmelakukanproses document fingerprinting. Proses iniditujukan agar dapatmengidentifikasipenjiplakan, termasukbagian-bagiankecil yang miripdalamdokumen yang berjumlahbanyak. Input dari proses document finger-printing adalah file teks. Kemudianoutputnyaakanberupasekumpulannilaihash yang disebutfingerprint. Fingerprint inilah yang akandijadikandasarpembandingantara file-file teks yang telahdimasukkan [4].

AnalisisMetode

Analisis metode ini merupakan penjelasan mengenai tahap-tahap pada aplikasi pendeteksian dokumen similaritas. Tahap pertama yang dilakukan adalah copy atau upload dokumen 1 dan 2 ke form input yang ada di website, tahap kedua yaitu mengecek kemiripan dokumen, kemudian tahap ketiga adalah menampilkan persentase kemiripan dan selesai. PadaGambar 1 berikutditunjukkan diagram aluruntukpenelitianini.



Gambar 1. Aplikasi Pendeteksian Dokumen Similaritas

Padaalgoritma winnowing ini terdapat beberapa tahap yaitu tahap pertama penghapusan karakter-karakter yang tidak relevan (white-space insensitivity), antara lain spasi atau tanda baca. Kemudian tahap kedua yaitu pembentukan rangkaian gram dengan ukuran k. Tahap ketiga adalah perhitungan nilai hash menggunakan rolling hash kedalam finger-printing dan tahap terakhir adalah menentukan persentase kesamaan antara 2 dokumen dengan persamaan Jaccard Coefficient.

Berikut ini adalah contoh kasus dan cara manual untuk menghitung nilai similaritas menggunakan algoritma winnowing :

Contoh kasus

Teks 1 : bunga mawar merah

Teks 2 : bunga mawar putih

1. Langkahpertama : menghilangkantandabacadanspasi
   Teks 1 : bungamawarmerah
   Teks 2 : bungamawarputih
2. Langkahkedua :

pembentukanrangkaiannilai k-gram denganukuran 5.
Teks1 :bungaungamngamagamawamawamawarawarmwarmearmerrmeramerah
Teks2 :bungaungamngamagamawamawamawarawarpwarpuarputrputiputih

3. Langkahketiga : melakukanperhitungannilai-nilai hash darisetiap gram menggunakan rolling hash. Formula untukmenghitungnilai hash dapatdilihatpadapersamaan (1) berikut.

$$c1 * b^{(k-1)} + c2 * b^{(k-2)} + \cdots + c_{(k-1)} * b + ck \quad (1)$$

Menghitungnilai hash dari kata "bunga" :
$b * 11^4 + u * 11^3 + n * 11^2 + g * 11^1 + a * 11^0 =$
$98 * 14641 + 117 * 1331 + 110 * 121 + 103 * 11 + 97 * 1 =$
1434818 +155727 + 13310 + 1133 + 97= 1605085

Setelah mendapatkan nilai hash dari kata "bunga" maka untuk mencari nilai hash kata kedua yaitu "ungam" tidak perlu menggunakan rumus 1 lagi, karena pada kata kedua terdapat juga karakter pada kata pertama sehingga menggunakan rumus kedua untuk mencari nilai hash pada kata kedua dan seterusnya yang dapatdilihatpadapersamaan (2) berikut.

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

Hc_2 = (1605085 - c_1 * 11^4) * 11 + c_{10}= (1605085 - 98 * 14641) * 11 + 109= (1605085 – 1434818) * 11 + 109= 170267 * 11 + 109
= 1873046

Dengan begini tidak perlu melakukan iterasi dari indeks pertama sampai terakhir untuk menghitung nilai hash untuk gram ke-2 sampai terakhir.Hal ini tentu dapat menghemat biaya komputasi saat menghitung nilai hash dari sebuah gram. Hasil perhitungan nilai-nilai hash dari setiap gram menggunakan rolling hash Teks 1:
1605085, 1873046, 1760636, 1651505, 1578399, 1740556, 1591666, 1886480, 1586325, 1827725, 1745265.

4. Langkahkeempat :membentuk window darinilai-nilai hash denganukuran 4.
(1605085 1873046 1760636 1651505),
(1873046 1760636 1651505 1578399),
(1760636 1651505 1578399 1740556),
(1651505 1578399 1740556 1591666),
(1578399 1740556 1591666 1886480),
(1740556 1591666 1886480 1586325),
(1591666 1886480 1586325 1827725),
(1886480 1586325 1827725 1745265).
5. Langkahkelima :memilihnilai hash terkecildarisetiap widow untukdijadikansebagai fingerprint, bilaterdapatnilai minimum hash yang samamakahanyaditulissatusajayaitu nilai hash yang paling kecilpada window yang pertamaditemukannilai hash tersebut.

Hasil fingerprint Teks1 : [1605085,0] [1578399,4]

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

cil na da lai

hash-nya tidak perlu dituliskan lagi. Nilai 0, 4 dan 8 merupakan nilai indeks dari hash yang terbentuk oleh k-gram. Parameter nilai pada window digunakan untuk mengam-bil perwakilan nilai hash sebagai bagian fingerprint yang tepat.

Untuk perhitungan teks 2 sama seperti teks 1, berikut ini adalah nilai-nilai hash dan pembentukan window dari teks 2 :

Nilai hash Teks 2:
1605085 1873046 1760636 1651505 1578399 1740556 1591669 1886529 1586866 1833684 1810814

NilaiWindowTeks 2 denganukuran4 :
(1605085 1873046 1760636 1651505)
(1873046 1760636 1651505 1578399)
(1760636 1651505 1578399 1740556)
(1651505 1578399 1740556 1591669)
(1578399 1740556 1591669 1886529)
(1740556 1591669 1886529 1586866)
(1591669 1886529 1586866 1833684)
(1886529 1586866 1833684 1810814)

Hasil fingerprint teks2 : [1605085,0] [1578399,4] [1586866,8]

6. Langkahkeenam :pengukurannilaisimilaritas. Nilaisimilaritasditentukandengan formula padapersamaan (3) berikut.

$$S = \frac{2}{4} \text{ x } 100\% = 50\%$$

Jadihasildaritingkatkesamaanteks 1 danteks 2 adalah 50%

Use case Diagram

Padagambar*Use case* untuk *user* ini terdapatlangkah-langkah yang akandijelaskan, antara lain langkahpertamaadalah user dapatmeng-upload dokumen, kemudian user input k-gram dan window. Setelahitudokumen di cekdanmenampilkanhasilpresentaseny a.Use case diagram Upload Files dapat dilihat pada Gambar 2 berikut.

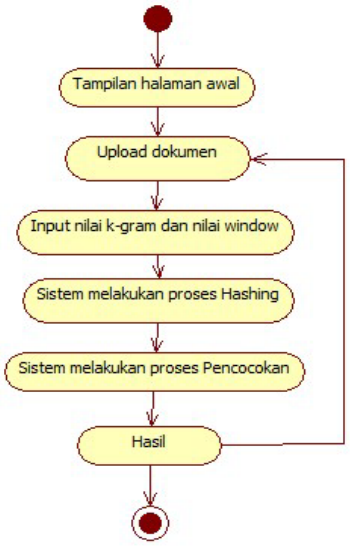
Gambar 2.*Use case diagram Upload Files*

Sedangkan untuk use case diagram Text only langkah awal dimu-lai dari user mengcopy dokumen yang berekstensi .txt, tahapkedua input *k-gram* dan *window* lalu kita bisa melihat hasil kesamaan dokumen tersebut berupa persentase. Use case diagram Text Only dapat dilihat pada Gambar 3 berikut.

Gambar 3.*Use case diagram Text Only*

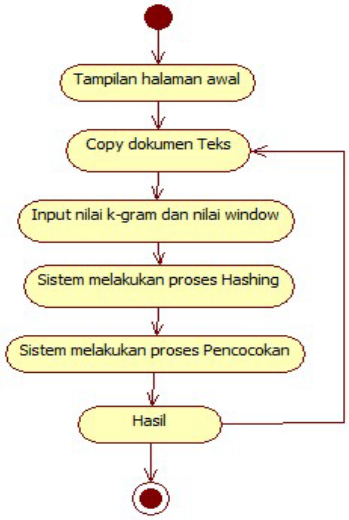
Activity Diagram

Langkah pertamadari activity diagram inidimulai dengan mengakses halaman awal pada website. Kemudian *user* dapat menuju ke halaman selan-jutnya yaitu halaman upload files, *user* dapat meng-*upload* 2. Lalu dokumen tersebut akan di proses. Setelah itu akan terlihat hasilnyadan akan menunjukan berapa persentasenya. Activity Diagram Upload Files dapat dilihat pada Gambar 4 berikut.



Gambar 4.*Activity Diagram Upload Files*

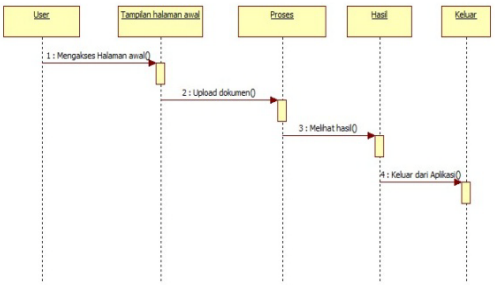
Sedangkanuntuk activity diagram Text Only adalah*user*dapatmeng-*copy* 2 dokumen. Laludokumentersebutakan di proses. Setelahituakanterlihathasilnyadanakan menunjukanberapapersentasenya. Activity diagram Text Only dapatdilihatpadaGambar5 berikut.



Gambar 5.*Activity Diagram Text Only*

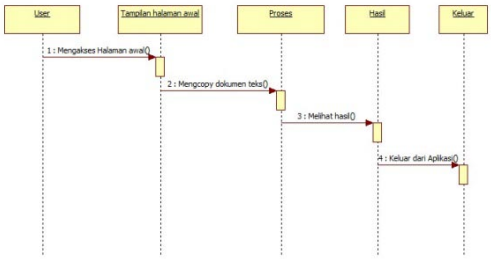
Sequence Diagram

Tahappertamadari sequence diagram adalah*User*dapatmengakseshalamanup load files untukmeng-*upload*dokumendanmelihathasildari*fin gerprinting*lalukeluardariaplikasi.Sequ ence diagram upload files dapatdilihatpadaGambar6 berikut.



Gambar 6.*Sequemce diagram Upload Files*

Sedangkan*Sequence diagram*untuk Text Only adalah*User*dapatmengakseshalamanTe xt Only untukmeng-*copy* dokumendanmelihathasildari*fingerprin ting*lalukeluardariaplikasi. Sequence diagram text only dapatdilihatpadaGambar7 berikut.

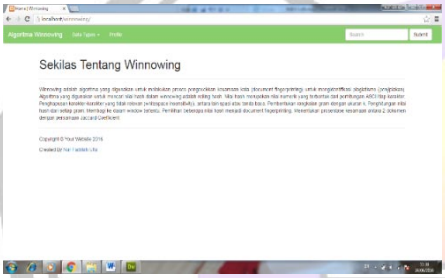


Gambar 7.*Sequemce diagram Text Only*

HASIL DAN PEMBAHASAN

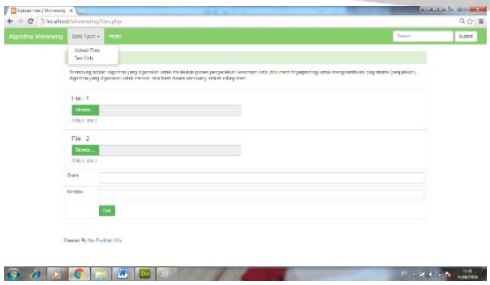
Tampilan Aplikasi

Tampilan awal terdiri dari sekilas tentang pengertin Algoritma Winnowing. Halamanawal website dapat dilihat pada Gambar 8 berikut.



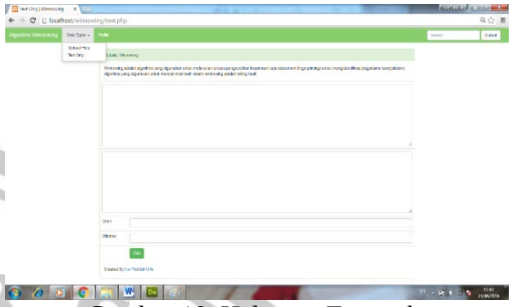
Gambar 8. Halaman Awal Website

Setelah pembuatan tampilan awal website, selanjutnya membuat tampilan untuk upload dokumen dan text only. Pada tampilan kedua ini terdapat penjelasan winnowing, kemudian adatulisan file-1 dan file-2serta tombol browse Selanjutnyaada gram dan juga window. Lalu tombol Cek. Halaman upload files dapat dilihat pada Gambar 9 berikut.



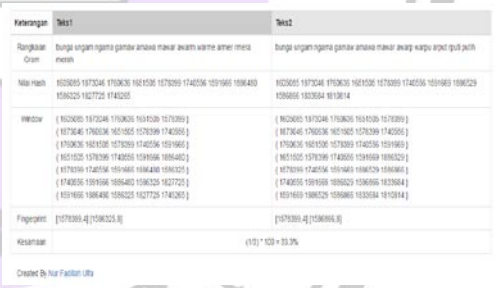
Gambar 9. Halaman Upload files

Pada tampilan di Text only ada kolom kosong yang akan di isi teks. Selanjutnyaadagram dan window.Laluada juga tombol Cek. Halaman text only dapat dilihat pada Gambar 10 berikut.



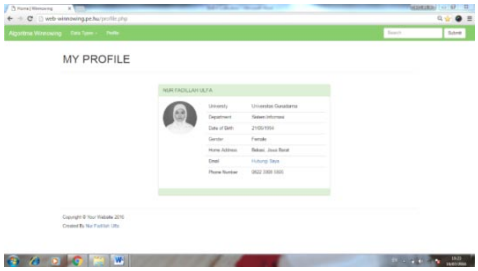
Gambar 10. Halaman *Text only*

Tampilan selanjutnya adalah tampilan keterangan teks 1 dan teks 2 yang sudah di uji. Tampilan keterangan ini dibuat menggunakan table yaitu table table-bordered. Halamanketerangan dapat dilihat pada Gambar 11 berikut.



Gambar 11. Halaman Keterangan

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. Halaman my profile dapat dilihat pada Gambar 12 berikut.



Gambar 12. Halaman *My Profile*.

PerbandinganEksekusiDokumendan Waktu

Berikutiniadalahtabelhasilujicoba perbandinganukurandoku-men danberapa lama waktu yang digunakan.

Tabel 1PerbandinganEksekusiDokumendanWaktu

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

KESIMPULAN DAN SARAN

Dari hasil ujicoba yang telah dilakukan yaitu pengecekan kemiripan dokumen menggunakan Algoritma Winnowing maka dapat diambil ke-simpulan bahwa aplikasi website ini akan memberikan hasil berupa persentase dan keterangan bahwa kedua dokumen yang diuji termasuk plagiat atau tidak.Aplikasi ini memiliki 5 kategori persentase kemi-ripannya sesuai dengan teori yang ada yaitu tidak plagiat, sedikit kesamaan, plagiat tingkat sedang, mendekati plagiarism dan plagiarism.Waktu proses untuk pendeteksian ini lebih lama jika memproses file yang cukup besar.

DAFTAR PUSTAKA

[1] Ridhatillah, Ardini 2003, Dealing with Plagiarism in the Information System Research Community: A Look at F Actors That Drive Plagiarism and Ways to Address Them, MIS Quarterly; Vol. 27, No. 4, pp. 511-532.

[2] Kurniawati, A., Sekarwati, Kemal A., danWicaksana, I wayanSimri. 2012. ArsitekturUntukAplikasiDeteksi KesamaanDokumen Bahasa Indonesia . Konferensi Nasional SistemInformasi 2012, STMIK - STIKOM Bali 23-25 Pebruari 2012. pp. 297-302.

[3] RiyanPratama, Mudafiq.2013PenerapanTeknik Document Fingerprinting PadaSistemPendeteksiPlagiarism eDokumenteksTerkelompokMen ggunakanAlgoritmaWinnowing DenganMetode K-Gram.JurusanTeknikInformatika UniversitasMuhammadiyah Malang.

[4] Schleimer, Saul, Wilkerson, Daniel s., Aiken, Alex. 2003. Winnowing: Local Algorithms for Document Fingerprinting. International conference on management of data. Proceedings of the 2003 ACM SIGMOD international conference on Management of data . pp. 76-85