PERANGKAT LUNAK PENDETEKSI KEMIRIPAN FILE DENGAN METODE TEXT MINING BERBASIS WEB

Rikky Wisnu Nugraha[1] , Astri Budiarti[2]

Konsentrasi Teknik Informatika, Program Studi Teknik Informatika, STMIKLPKIA
Jln. Soekarno Hatta No. 456 Bandung 40266, Telp. +62 22 75642823, Fax. +62 22 7564282
Email :r.wisnunugraha@gmail.com[1],as3budiarti@yahoo.com[2]

Abstrak

Perkembangan teknologi dan informasi saat ini sangatlah pesat, termasuk pamanfaatan internet, dimana dengan internet kita dapat dengan mudah memperoleh informasi kapanpun dan dimanapun. Dengan kemudahan tersebut sangat mungkin terjadinya plagiarisme, tidak menutup kemungkinan plagiarisme juga terjadi di dunia pendidikan.
Plagiarisme merupakan tindakan mengambil hasil pemikiran orang lain dan diakui sebagai hasil pemikiran sendiri. Plagirisme ini semakin berkembang seiring kemajuan teknologi. Guna meminimalisir hal tersebut digunakan metode text mining dalam pendeteksiannya juga metode *string matching* dalam perhitungan kemiripan persentasenya.
Perangkat lunak pendeteksi kemiripan file ini menggunakan algortima *stoplist* dan algoritma Nazief dan Adriani dalam pemrosesan teks nya sebagai pendukung dari metode text mining.
Perangkat lunak pendeteksi kemiripan file ini dirancang dengan pemodelan terstruktur dengan menggunakan *process hierarcy diagram, bussiness process diagram* dan *process specification* serta dibangun dengan menggunakan bahasa pemograman php dengan database mysql.
Perangkat lunak yang dibangun diharapkan akan meminimalisir plagiarisme. Perangkat lunak ini sudah diuji oleh 20 orang user melalui kuisioner, dan dapat ditarik kesimpulan bahwa dengan dibangunnya perangkat lunak ini, dapat memberi kemudahan dalam proses pendeteksian kemiripan file, serta menyediakan informasi dan fasilitas yang lengkap untuk user.

Kata kunci : *plagiarisme, text mining, algortima stoplist, algortima nazief dan adriani*

1. Pendahuluan

Kemudahan pertukaran informasi merupakan salah satu dampak positif dari kemajuan teknologi yang semakin pesat pada saat ini, namun disamping itu terdapat pula dampak negatif nya yaitu semakin maraknya plagiarisme.

Plagiarisme adalah suatu tindakan menjiplak hasil karya/hasil pemikiran orang lain sebagai hasil karya/hasil pemikiran sendiri. Tindakan Plagiarisme sangat buruk dampaknya, baik bagi si pemilik asli ataupun bagi si pelaku. Plagiarisme dapat mematikan kreatifitas seseorang karena sudah terbiasa mengambil hasil pemikiran orang lain yang diakuinya sebagi hasil pemikiran sendiri. Asal mula tindakan plagiarisme berawal dari sikap malas seseorang dan tidak mau berfikir keras untuk menghasilkan ide original, maka dari itu sifat plagiat harus dihilangkan sejak dini.

Plagiarisme sering terjadi khususnya di bidang akademis, sebagai contohnya mahasiswa menjiplak dokumen digital dari internet atau meng*copy-paste* hasil pekerjaan temannya sehingga menjadi suatu hasil yang berbeda dari sebelumnya. Guna meminimalisir hal tersebut banyak dilakukan deteksi plagiat secara manual, misalnya seorang dosen melakukan pengecekan terhadap gaya penulisan file mahasiswanya. Akan tetapi, deteksi manual memiliki masalah pada efisiensi waktu dan keakuratan hasilnya. Sangat tidak mungkin memeriksa puluhan file secara bersamaan sehingga harus dilakukan dalam beberapa tahap yang tentu membutuhkan waktu yang cukup lama.

Masalah-masalah yang dihadapi suatu lembaga pendidikan XYZ dapat di identifikasikan sebagai berikut:

1. User (tenaga pendidik) kesulitan memeriksa file dalam kuantitas banyak secara akurat.
2. Tidak objektifnya penilai suatu tugas
3. Tenaga pendidik sulit mengawasi mahasiswa yang sering melakukan penjiplakan.

Dalam penyusunan jurnal ini telah dibatasi ruang lingkup permasalahan hanya pada proses mendeteksi kemiripan file. Agar pembahasan tidak menyimpang dari maksud yang ingin disampaikan yang mencakup:

1. Mendeteksi file dengan type doc (Microsoft Word 2003) dan pdf
2. File berisi text berbahasa indonesia

3. File berisi text, bukan gambar
4. Tidak memperhatikan kesalahan ejaan/pengetikan
5. Memeriksa maksimal 30 file

Adapun tujuan dari perancangan sistem ini adalah sebagai berikut :

1. Membantu user mengetahui persentase kemiripan file-file yang dibandingkan
2. Meminimalisir human error saat pemeriksaan file dalam jumlah banyak
3. Menyediakan media pengawas originalitas hasil karya mahasiswa bagi user pemeriksa.

2. Landasan Teori

Rekayasa perangkat lunak yang akan dikembangkan, merupakan pembaruan perangkat lunak pendeteksi kemiripan yang sudah ada sebelumnya.

Menurut Rosa A.S dan M. Shalahuddin (2011) dalam bukunya "Modul Pembelajaran Rekayasa Perangkat Lunak" mendefinisikan bahwa perangkat lunak adalah "Perangkat lunak (software) adalah program komputer yang terasosiasi dengan dokumentasi perangkat lunak seperti dokumentasi kebutuhan, model desain, dan cara penggunaan (user manual)".

Karakter perangkat lunak adalah sebagai berikut:

1. Perangkat lunak dibangun dengan rekayasa (*software engineering*) bukan diproduksi secara manufaktur atau pabrikan.
2. Perangkat lunak tidak pernah usang ("*wear out*") karena kecatatan dalam perangkat lunak dapat diperbaiki.
3. Barang produksi pabrikan biasanya komponen barunya akan terus diproduksi, sedangkan perangkat lunak biasanya terus diperbaiki seiring bertambahnya kebutuhan.

Menurut Indrajani (2011) dalam bukunya "Pengantar dan Sistem Basis Data" mendefinisikan bahwa web adalah : "Halaman-halaman yang digunakan untuk menampilkan informasi, gambar gerak, suara, dan atau gabungan dari semuanya itu, baik yang bersifat statis maupun dinamis yang membentuk satu rangkaian bangunan yang saling berhubungan melalui link-link".

Menurut Kamus Besar Bahasa Indonesia (KBBI), mendefinisikan Plagiarisme adalah : "*Tindakan penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri.*".

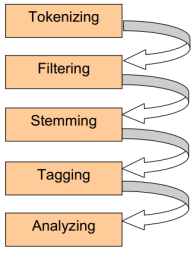
Menurut Mozgovoy (2006), sistem pendeteksi plagiarsime dapat dikelompokkan menjadi tiga kelompok, yaitu:

1. *Fingerprint Based*
   Fingerprint mengandung beberapa atribut dan merefleksikan struktur dari dokumen. Atribut tersebut diantaranya jumlah kata per baris, jumlah kata unik, dan jumlah kutipan pendek. Setiap dokumen nantinya akan dibuat *fingerprint*. Jika *fingerprint* antara dua dokumen mirip, maka bisa dikatakan salah satu merupakan plagiat.
2. *String-Matching Based*
   Metode *string-matching based* membandingkan dokumen dengan memandangnya sebagai strings. Masing-masing string dari satu dokumen akan dibandingkan dengan string dokumen lain dan dihitung kesamaannya.
3. *Tree-Matching Based*
   Dokumen yang dibandingkan harus memiliki aturan struktur yang sama. Metode ini sangat cocok untuk mendeteksi plagiat pada source program komputer karena terstruktur secara jelas.

Dari ketiga metode diatas, penyusun menggunaka metode *String-Matching Based*dalam proses pendeteksian file. Metode tersebut diimplementasikan pada perangkat lunak dengan menggunakan text mining. Untuk mendapatkan kata kunci dari suatu file, kata-kata dalam file diolah dengan menggunakan beberapa tahapan dari *text mining*, setelah itu dihitung banyaknya kemunculan string yang dianggap kata kunci lalu dibandingkan dengan kata kunci pada file lain dan dihitung persentase kemiripannya.

Menurut Even (2002) dalam jurnalnya *Text mining* merupakan "*Salah satu bentuk eksplorasi dan analisis data teks yang bertujuan untuk mendapatkan pengetahuan baru baik itu melalui cara otomatis maupun semi otomatis*".

Menurut Milkha Harlian dalam jurnalnya "*Text Mining*" terdapat 5 tahapan dalam tex mining yaitu :



Gambar 1 Tahapan *Text Mining*

Algoritma *stop list* digunakan dalam salah satu tahapan *text mining* yaitu *filtering*. *Stop list* adalah proses membuang atau menghilangkan kata-kata yang kurang penting, seperti kata depan, kata sambung, kata ganti, dll. (Daftar *Stop list* dilampirkan)
Algoritma ini merupakan algoritma *stemming* yang dikembangkan oleh Bobby Nazief dan Mirna Adriani pada tahun 1996 sebagai hasil penelitian internal Universitas Indonesia. Dengan menggabungkan metode pemotongan imbuhan serta pencarian kamus yang terdiri atas kata dasar, algoritma ini menghasilkan tingkat akurasi *stemming* pada teks bahasa Indonesia yang lebih tinggi dibandingkan algoritma Porter.
Proses *stemming* ini dilakukan dengan cara memotong imbuhan yang dilakukan secara rekursif. Serta mencari kata didalam kamus yang dilakukan sebelum tahap pemotongan. Kelemahan dari algoritma ini diantaranya tingkat akurasinya tergantung dari kamus yang dimiliki. Algoritma Stemming Bahasa Indonesia Nazief dan Adriani ini mempunyai aturan imbuhan sendiri dengan model, seperti :

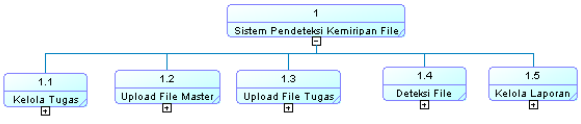
Prefiks 1 + Prefiks 2 + Kata dasar + Sufiks 3 + Sufiks 2 + Sufiks 1

Metodologi yang digunakan dalam pembuatan perangkat lunak ini menggunakan metodologi analisis dan terstruktur dengan pemodelan sekuensial linear yang mengambil pendapat dari Rosa A.S dan M. Shalahuddin (2010) dengan bukunya yang berjudul "*Modul Pembelajaran Rekayasa Perangkat Lunak*".

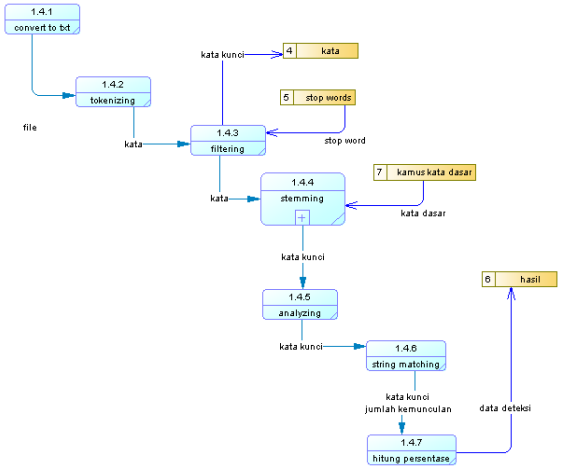
Sekuensial linear mengusulkan sebuah pendekatan kepada pengembangan perangkat lunak yang sistematik dan sekuensial yang dimulai pada tingkat dan kemajuan sistem pada seluruh analisis, desain, kode, pengujian, dan pemeliharaan.

3. Hasil Penelitian

Untuk mempermudah dalam merancang dan menggambarkan proses yang akan di rancang, maka dibuatkanlah *Business Process Diagram* yang akan menggambarkan secara terstruktur dan sederhana semua proses yang ada dalam sistem. Dalam *Business Process Diagram* terdapat *Process Hierarchy Diagram* yang merupakan gambaran sederhana ruang lingkup dan hubungannya dengan external entities.



Business Process Diagram di bawah ini menggambarkan semua proses dalam mendeteksi file, dimulai dari proses mengkonversi file ke bentuk txt, proses *text mining*, proses *string matching* hingga proses perhitungan persentasi kemiripan.



Dalam tahap implementasiini terdapat banyak kegiatan yang meliputi pembuatan perangkat lunak, maka berdasarkan kegiatan-kegiatan yang ada dipaparkan seluruh kegiatan seperti di bawah ini.

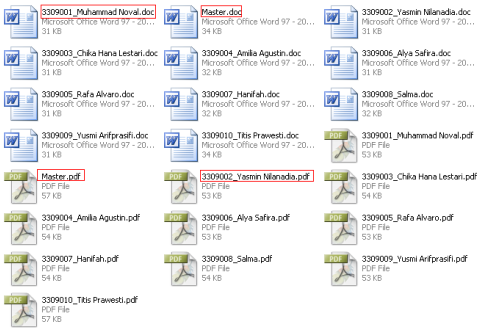
Kode Aktivitas	Keterangan Aktivitas	Waktu (minggu)	Predecessor
A	Analisa Kebutuhan Sistem	2	-
B	Pengumpulan Data	2	A
C	Perancangan Sistem	6	A,B
D	Pemrograman	4	C
E	Bimbingan Keseluruhan Dan Koreksi	1	D
F	Menerapkan Program Komputer	1	E

Perancangan Antarmuka

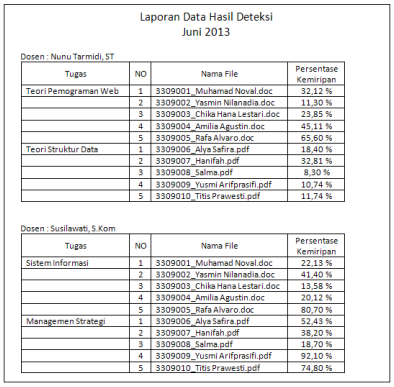
Perancangan antarmuka merupakan penjelasan mengenai bagaimana perangkat lunak dapat digunakan dengan sebaik mungkin oleh pengguna, agar setiap penggunan perangkat lunak ini dapat mempermudah penggunanya. Untuk itu dibuatkan struktur menu perangkat lunak pendeteksi kemiripan file yang selanjutnya dibuatkan dialog-dialog screen yang dapat digambarkan dan penjabaran perangkat lunak ini sehingga mempermudah pengguna menggunakan dan melihat tampilan perangkat lunak yang sebenarnya, berikut adalah gambaran dari struktur menu yang telah dibuat.

Form Input Output

Dibawah ini merupakan format input file :



Dan form output hasil deteksi :



Laporan Data Hasil Deteksi
Juni 2013

Dosen : Nunu Tarmidi, ST

Tugas	NO	Nama File	Persentase Kemiripan
Teori Pemograman Web	1	3309001_Muhamad Noval.doc	32,12 %
	2	3309002_Yasmin Nilanadia.doc	11,30 %
	3	3309003_Chika Hana Lestari.doc	23,85 %
	4	3309004_Amilia Agustin.doc	45,11 %
	5	3309005_Rafa Alvaro.doc	65,60 %
Teori Struktur Data	1	3309006_Alya Safira.pdf	18,40 %
	2	3309007_Hanifah.pdf	32,81 %
	3	3309008_Salma.pdf	8,30 %
	4	3309009_Yusmi Arifprasifi.pdf	10,74 %
	5	3309010_Titis Prawesti.pdf	11,74 %

Dosen : Susilawati, S.Kom

Tugas	NO	Nama File	Persentase Kemiripan
Sistem Informasi	1	3309001_Muhamad Noval.doc	22,13 %
	2	3309002_Yasmin Nilanadia.doc	41,40 %
	3	3309003_Chika Hana Lestari.doc	13,58 %
	4	3309004_Amilia Agustin.doc	20,12 %
	5	3309005_Rafa Alvaro.doc	80,70 %
Managemen Strategi	1	3309006_Alya Safira.pdf	52,43 %
	2	3309007_Hanifah.pdf	38,20 %
	3	3309008_Salma.pdf	18,70 %
	4	3309009_Yusmi Arifprasifi.pdf	92,10 %
	5	3309010_Titis Prawesti.pdf	74,80 %

Untuk menjalankan sebuah program, dibutuhkan perangkat pendukung yang dapat dijalankan dengan baik. Untuk itu dibutuhkan *software* dan *hardware* yang benar – benar support dengan perangkat yang dibutuhkan untuk membuat program aplikasi serta manusia yang terlibat dalam hubungannya dengan perangkat lunak (*brainware*). *Spesifikasi hardware* dan *software* diantaranya :

Kebutuhan Perangkat Lunak (*Software*)

1. Sistem Operasi minimum Windows SP3.
2. Web Server yang digunakan adalah Apache/2.2.8 (Win32)
3. Aplikasi Server menggunakan XAMPP Versi 1.6.6a
4. Bahasa *Scripting* yang digunakan adalah PHP Versi 5.2.5
5. Software database yang digunakan adalah MySQL phpmyadmin Versi 5.0.51a
6. Web Browser menggunakan Mozilla Firefox 12.0
7. *Plugin*s antiword dan xpdf

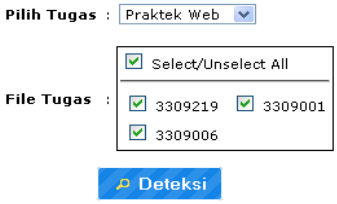
Kebutuhan Perangkat Keras (*Hardware*)

1. Koneksi LAN
2. Personal Computer / Laptop
3. Printer untuk mencetak laporan

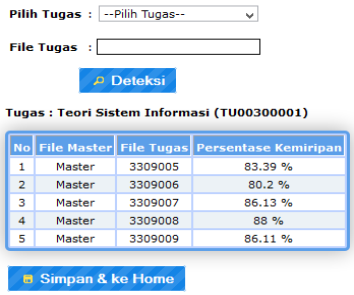
Kebutuhan Brainware

1. User(dosen)
   Orang yang melakukan deteksi file pada perangkat lunak.
2. Administrator
   Orang yang mengoperasikan atau mengelola perangkat lunak pendeteksi kemiripan file dengan hak akses sebagai administrator.

Berikut dialog screen saat melakukan deteksi :



Dan berikut adalah dialog screen hasil deteksi :



Pengujian

Pengujian merupakan bagian yang penting dalam siklus pembangunan perangkat lunak.Pengujian dilakukan untuk menjamin kualitas dan juga mengetahui kelemahan dari perangkat lunak. Tujuan dari pengujian ini adalah untuk menjamin bahwa perangkat lunak yang dibangun memiliki kualitas yang handal, yaitu merepresentasikan kajian pokok dari spesifikasi, analis perancangan dan pengkodean dari perangkat lunak itu sendiri.

Pada tahap ini dilakukan pengujian terhadap modul program yang menghasilkan pengujian yang telah dilakukan. Pada tahap pengujian ini, test pengujian menggunakan metode test pengujian *black box.*
Pengujian *black box* berfokus pada persyaratan fungsional perangkat lunak. Dengan demikian,

pengujian *black-box* memungkinkan perekayasa perangkat lunak mendapatkan serangkaian kondisi input yang sepenuhnya menggunakan semua persyaratan fungsional untuk suatu program. Berikut adalah tabel pengujian yang telah dibuat.

Kesimpulan Pengujian

Berdasarkan hasil pengujian dengan kasus uji sample diatas(kuesioner terlamir) dapat ditarik kesimpulan bahwa perangkat lunak secara fungsional mengeluarkan hasil yang sesuai dengan yang diharapkan dan tampilan yang menarik serta fasilitas yang cukup lengkap pada perangkat lunak.

4. Kesimpulan

Setelah melakukan penelitian terhadap perangkat lunak yang dibuat dan mengimplementasikan serta pengujian yang dilakukan, berikut ini adalah hasil kesimpulan yang dapat dijabarkan:

1. User (tenaga pendidik) mendapatkan kemudahan dalam memeriksa file tugas dengan jumlah banyak melalui perangkat lunak berbasis web juga client server sehingga waktu proses pemeriksaan menjadi cepat dan akurat.
2. Terdapat bukti akurat terhadap suatu file dinyatakan hasil penjiplakan dengan adanya persentase kemiripan pada hasil deteksi di perangkat lunak.
3. Tenaga pendidik mudah melakukan pengawasan terhadap mahasiswa yang sering melakukan penjiplakan melalui laporan yang dihasilkan perangkat lunak sehingga dapat memberikan sanksi pada mahasiswa yang sering melakukan penjiplakan dalam pengerjaan tugas.

DAFTAR PUSTAKA

1. Abdul, Kadir. 2008, *Dasar Pemograman Web Dinamis Menggunakan PHP*, Edisi III, Andi, Yogyakarta.
2. Abdul, Kadir. 2000, *Mastering Ajax dan PHP*, Edisi I, Andi, Yogyakarta.
3. Eko, Nugroho. 2011, *Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp*, Universitas Brawijaya, Malang.
4. Feldman, Ronen dan James Sanger. 2007, *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press.
5. Kamus Besar Bahasa Indonesia, http://kamusbahasaindonesia.org/.
6. Milkha Harlian. 2006, *Text Mining*, University of Texas, Austin.
7. Mozgovoy, Maxim. 2007, *Enhancing Computer-Aided Plagiarism Detection*, University Of Joensuu, 2007.
8. Mudafiq Riyan, et all, *Aplikasi Pendeteksi Duplikasi Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Winnowing Dengan Metode K-Gram Dan Synonym Recognition*, Universitas Muhammadiyah, Malang.