1. Pendahuluan

1.1 Latar belakang

Plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain [1]. Beberapa tipe plagiarisme yaitu : [8]

a. Word-for-word plagiarism
   Menyalin setiap kata secara langsung tanpa diubah sedikitpun.
b. Plagirism of authorship
   Mengakui hasil karya orang lain sebagai hasil karya sendiri dengan cara mencantumkan nama sendiri menggantikan nama pengarang yang sebenarnya.
c. Plagiarism of ideas
   Mengakui hasil pemikiran atau ide orang lain.
d. Plagiarism of sources
   Jika seorang penulis menggunakan kutipan dari penulis lainnya tanpa mencantumkan sumbernya.

Akan sangat sulit mengukur kesamaan suatu dokumen hanya dengan membaca sekali saja, pembaca diharuskan membaca berulang-ulang untuk memastikan apakah dokumen tersebut sama atau tidak. Akan lebih sulit lagi apabila yang harus dibandingkan lebih dari dua dokumen. Oleh karena itu, dibutuhkan suatu sistem atau aplikasi yang dapat mendeteksi kesamaan dokumen secara otomatis dan akurat.

Sebelum dokumen dapat di analisis menggunakan algoritma RKR-GST, maka perlu dilakukan *prepocessing* terhadap dokumen tersebut dengan menggunakan *Text Mining*. *Text Mining* bertujuan untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapa dilakukan analisis keterhubungan antar dokumen. Adapun tahapan *Text Mining* yang dilakukan secara umum adalah:[3]

a. *Tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
b. *Filtering* adalah tahap mengambil kata-kata penting dari hasil token.
c. *Stemming* adalah tahap mencari root kata dari tiap kata hasil filtering.
d. *Tagging* adalah tahap mencari bentuk awal/root dari tiap kata lampau atau kata hasil stemming.
e. *Analizing* adalah tahap penentuan seberapa jauh keterhubungan antar kata-kata antar dokumen yang ada.

Alasan mengapa algoritma RKR-GST dipilih untuk kasus pendeteksian dokumen karena : [11]

a. Algoritma RKR-GST dapat diterapkan pada *'free'* teks, seperti esai, artikel surat kabar dan teks lainnya .
b. Algoritma RKR-GST dapat mengidentifikasi secara acak penyisipan atau penghapusan teks.
c. Dalam eksperimen diperoleh rata-rata kompleksitas, [18], mendekati garis lurus, yang membuatnya menjadi pilihan yang lebih baik daripada Algortima GST. Jadi algoritma dapat diterapkan pada teks-teks yang relatif panjang.

d. Algoritma dapat menangani kasus-kasus dokumen yang berisi nama, tanggal, lokasi, domain spesifik kata-kata atau istilah pengetahuan umum dalam jumlah besar. Biasanya kata-kata ini akan menjadi bagian dari pencocokan string yang cukup panjang dan tidak hanya berdiri sendiri.
e. Tiga term baru diperkenalkan dalam definisi RKR-GST yaitu parameter s*earch-length*, *hash-value* dan *Karp-Rabin (KR) algorithm*. Ini ditambahkan dari term GST yaitu *Maximal-match* , *Tile* dan *Minimum-match-length* .

1.2 Perumusan masalah

Permasalahan yang akan diangkat dalam tugas akhir ini adalah:

a. Bagaimana melakukan preprocessing dari dokumen sebelum di identifikasi kesamaannya.
b. Bagaimana menganalisis Algoritma RKR-GST dalam pendeteksian kesamaan dokumen.
c. Bagaimana melakukan proses pendeteksian kesamaan dokumen berdasarkan kesamaan kata dan atau perubahannya menggunakan algoritma RKR-GST.

Batasan Masalah

Adapun batasan dalam tugas akhir ini adalah:

a. Tipe dokumen yang dibandingkan berupa file text.
b. Parameter pembobotan term yang digunakan adalah s*earch-length*, *hash-value* dan *Karp-Rabin (KR) algorithm*.
c. Hasil yang diberikan oleh sistem sebagai parameter akurasi adalah MML (*Minimum Match Length*) value, persentase nilai kesamaan (*similarity value*) disertai dengan representasi visual.
d. Dokumen yang dibandingkan bukan merupakan dokumen ringkasan, tetapi sesuai dengan template/ satu ruang lingkup permasalahan dan di inputkan ke dalam sistem.
e. Tidak membandingkan dengan algoritma lain yang dapat digunakan untuk mendeteksi plagiarisme.
f. Hanya sebagai alat pembantu mendeteksi kesamaan dokumen. Keputusan dokumen tersebut plagiat atau bukan, tergantung kepada pembaca atau user.
g. Kategori plagiat dokumen pengujian berdasarkan tingkat kesamaannya dibagi menjadi *heavily*,*lightly plagiarized* dan *original text*.

1.3 Tujuan

Yang menjadi tujuan dibuatnya tugas akhir ini antara lain:

a. Membangun suatu aplikasi yang mengimplementasikan algoritma RKR-GST untuk pendeteksian kesamaan dokumen, dengan menerapkan *tokenizing* sebagai tahapan prepocessing.
b. Mengidentifikasi suatu dokumen dengan dokumen lain berdasarkan tingkat kesamaan kata dan atau perubahannya menggunakan algoritma RKR-GST untuk mengetahui seberapa besar kesamaannya.

c. Menganalisis sistem pendeteksian kesamaan dokumen yang telah dibangun.

Hipotesis yang akan dibuktikan dari penelitian Tugas Akhir ini adalah :
Dengan menggunakan algoritma RKR-GST, dokumen yang diuji dapat dipisahkan kedalam katogeri plagiat dengan baik berdasarkan threshold / similarity value level yang dihasilkan oleh algoritma tersebut. Dimana yang dikategorikan sebagai heavily plagiarized, persentase nilai kesamaannya (similarity) seharusnya adalah diatas 50%.(berdasarkan penelitian Wilk tahun 2002[11])

1.4 Metodologi penyelesaian masalah

Untuk pembangunan sistem digunakan metode sebagai berikut:

a. Studi Literatur
   1. Mengumpulkan dan mempelajari referensi berupa makalah, jurnal, e-book maupun buku mengenai Text Mining dan Algoritma RKR-GST yang berkaitan dengan Tugas Akhir .

b. Observasi
   1. Melakukan diskusi dan pembahasan, baik dengan pembimbing maupun dengan orang yang berkompeten mengenai Text Mining, Algoritma RKR-GST dan kasus pendeteksian kesamaan dokumen.

c. Analisis dan design
   Melakukan analisis dari identifikasi masalah dan studi literatur yang dilakukan untuk menentukan solusi dari perumusan masalah kemudian membangun desain dari program yang akan diterapkan. Yang dianalisis adalah:
   1. Menganalisis data, fungsionalitas dan behavioral requirement yang akan digunakan dalam pendeteksian kesamaan dokumen.
   2. Menganalisis Prepocessing yang akan digunakan dan Algoritma RKR-GST yang digunakan dalam pendeteksian kesamaan dokumen.

d. Implementasi
   Adapun tahapan dalam implementasi Algoritma RKR-GST dalam kasus pendeteksian kesamaan dokumen untuk kasus ini yaitu:
   1. Tokenizing, adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
   2. Implementasi algoritma RKR-GST menggunakan bahasa pemrograman.
   3. Evaluasi statistik untuk mentukan tingkat kesamaan dokumen.

e. Test
   Melakukan pengujian terhadap sistem yang dibangun dengan cara melakukan pengelompokan dokumen uji yang terdiri dari:
   1. ‘Heavily’ plagiarised text yaitu menduplikasi kalimat dari *document resource* sehingga terlihat sama dengan dokumen aslinya.
   2. ‘Lightly’ plagiarized text yaitu menuliskan kembali dokumen yang kalimatnya kebanyakan adalah kalimat sendiri (tidak sama dengan dokumen aslinya).

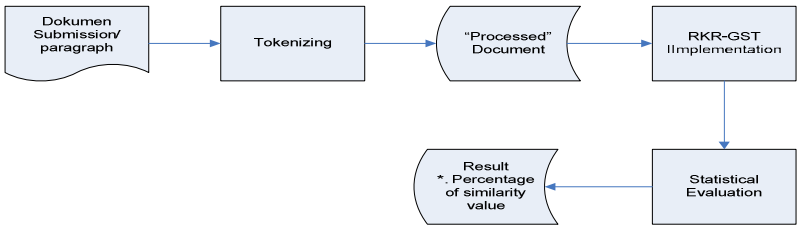
3. ‘Rewritten’ text yaitu membuat dokumen dengan kata-kata sendiri atau bisa dikategorikan sebagai original text.

Jumlah dokumen uji untuk masing-masing kategori akan disesuaian dengan kebutuhan system sehingga dapat menghasilkan hasil yang optimal.

f. Pembuatan laporan

Menyusun laporan tertulis berdasarkan hasil penelitian yang dilakukan dan memberikan kesimpulan serta saran untuk pengembangan program yang mungkin akan dibangun atau dikembangkan dimasa yang akan datang.

Deskripsi sistem :



Gambar 1-1 Deskripsi Sistem

Input adalah dokumen text atau paragraph.

Output:

a. Dokumen yang siap di proses menggunakan algoritma RKR-GST.
b. Persentasi dari kesamaan dokumen.